

REPUBLIC OF TURKEY
ÇANAKKALE ONSEKİZ MART UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAMME



DEVELOPING A VALID AND RELIABLE SPEAKING TEST

MASTER'S THESIS

VAKKAS SELİM YÜKSEL

ÇANAKKALE

2019

Republic of Turkey
Çanakkale Onsekiz Mart University
Graduate School of Educational Sciences
Department of Foreign Languages Education
English Language Teaching Programme

Developing a Valid and Reliable Speaking Test

Vakkas Selim YÜKSEL
(Master's Thesis)

Supervisor
Assoc. Prof. Dr. Salim RAZI
Çanakkale

2019

Taahhütname

Yüksek Lisans tezi olarak sunduğum “Developing a Valid and Reliable Speaking Test” adlı çalışmanın, tarafımdan, bilimsel ahlak ve değerlere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yaralandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yaparak yararlanmış olduğumu belirtir ve bunu onurumla doğrularım.

28/08/2019

Vakkas Selim YÜKSEL



Çanakkale Onsekiz Mart Üniversitesi

Eğitim Bilimleri Enstitüsü

Onay

Vakkas Selim YÜKSEL tarafından hazırlanan çalışma, 28/08/2019 tarihinde yapılan tez savunma sonucunda jüri tarafından başarılı bulunmuş ve Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Referans No : 10288008

Akademik Unvan	Adı SOYADI	İmza
Doç. Dr.	Salim RAZI Danışman
Prof. Dr.	Dinçay KÖKSAL Üye
Prof. Dr.	Arif SARIÇOBAN Üye

Tarih:.....

İmza:.....

Prof. Dr. Salih Zeki GENÇ

Enstitü Müdürü

Acknowledgement

There are some people whom I owe debt of gratitude in the writing process of this thesis dissertation. First and the foremost, I would like to express my thankfulness to my mother and father for their endless love and support both throughout my life and this thesis dissertation. Secondly, I would like to thank to my elder brother and sister, for showing support when I feel weak and powerless.

Next, without doubt, to my wife, Mediha, I would like to thank so many times. Not only did she enlighten my life but also lighted the fire of motivation to write and complete this study. Without her, I wouldn't resume this study. Also, to my daughter, Defne, who is still in her mother's womb during which these lines are written and who has also enlarged our happiness.

I would also like to thank to my respected supervisor Assoc. Prof. Dr. Salim RAZI, who kept believing in me and helped a lot during the breakdowns in the writing process. Also, I want to express my gratitude to Prof. Dr. Dinçay KÖKSAL and Assoc. Prof. Dr. Ece ZEHİR TOPKAYA for teaching invaluable theoretical knowledge that I will always use in my career.

Last but not the least, I would like to thank to my administrator for giving me permission to go to Çanakkale to attend classes, to my colleagues being always guide in front of me, and students for being in my life. They have become a very big family for me. Finally, thanks to elder brother Süleyman ÜNAL, for giving this thesis subject as an idea and supporting me in writing this dissertation.

Çanakkale, 2019

Vakkas Selim YÜKSEL

Abstract

Developing A Valid and Reliable Speaking Test

Testing foreign language speaking abilities has been a problematic issue. Deciding on tasks, eliciting desired samples of speech, aspects of speech to assess, standards and procedures in making assessment are the challenges in speaking exam. Mistakes in these issues could create fractures in the two indispensable conditions of a test; validity and reliability. In this context, in preparation class of an engineering faculty also graduating pilot candidates, the necessity to a valid and reliable speaking test arose. Therefore, this study aimed to develop a valid and reliable speaking examination testing CEFR B1 proficiency level, the requirement of the institution. There are 164 test takers, 36 raters and 3 administrators as participants. To develop the test, firstly test specifications were defined, tasks were chosen and assessment scale was prepared. Then, a rater training session was held for 36 raters. Both quantitative and qualitative research methods were used in the study. Construct validity, inter-rater and intra-rater reliability were investigated through using statistical analysis software of SPSS 22.0 and standard error of measurement was investigated through an online statistical calculator. Content validity was investigated by comparing the abilities and topics in the speaking exam with objectives and content of the course book curriculum. An interview was conducted to find the views of administrators, teachers and students about the speaking exam. Investigation of construct validity through factor analysis revealed that nine items constituting the speaking score of a test taker were grouped under a factor with eigen-value of 6.42 explaining 72.1% of the variance. Inter-rater reliability analysis revealed that in 17 out of 18 commissions, raters' ratings correlated positively and strongly with each other ($r_s \geq .69$, $p. \leq .03$). Intra-rater reliability analysis revealed a

Cronbach's alpha score of $\alpha \geq .70$ over 9 items in the speaking test for 33 raters out of 36. Also, standard error of the speaking test was $SE = .095$. Investigation of content validity revealed that abilities tested in the speaking exam matched with the 70% of the abilities in the course book and 50% of the topics in the course book were tested in the exam. Results of interview revealed that the speaking exam was authentic, practical and created a positive washback effect on teachers and students. As a conclusion, the developed speaking test is a valid and a reliable test which also is authentic, practical and leaving positive washback effect.

Keywords: EFL speaking assessment, rating scale, rater training, reliability, sample speaking tasks, validity.

Özet

Geçerli ve Güvenilir bir Konuşma Sınavı Geliştirme

Yabancı dilde konuşma becerilerini test etme problematik bir konudur. Adayların konuşma becerilerini gerçekten yansıtan konuşma örneklerini ortaya çıkaran konuşma görevlerine, konuşmanın hangi yönlerinin değerlendirileceğine ve sınavı icra etmede ve değerlendirme de standartları sağlamaya karar vermek ve bunları tanımlamak konuşma sınavlarındaki güçlüklerdir. Bu konulardaki hatalar bir testin olmazsa olmaz iki koşulunda çatlaklara sebep olabilir; geçerlilik ve güvenilirlik. Bu bağlamda, aynı zamanda pilot aday yetiştiren mühendislik fakültesinin hazırlık sınıfında AODRÇ B1 yeterlilik seviyesini ölçebilecek geçerli ve güvenilir bir konuşma sınavı ihtiyacı ortaya çıkmıştır. Dolayısıyla, bu çalışmada kurumun gereksinimi olan CEFR B1 seviyesini ölçebilen geçerli ve güvenilir bir konuşma sınavı geliştirme amaçlanmıştır. Bu çalışmada 164 sınava girecek olan, 36 puanlayıcı, 3 yönetici vardır. Sınavı geliştirmek için öncelikle test özellikleri belirlenmiş, sınav görevleri seçilmiş ve değerlendirme ölçekleri hazırlanmıştır. Daha sonra 36 puanlayıcıya puanlama eğitimi verilmiştir. Bu çalışmada hem nicel hem de nitel araştırma yöntemleri kullanılmıştır. Yapı geçerliliğini, puanlayıcılar arası güvenilirliği, puanlayıcı iç güvenilirliğini araştırmak için SPSS 22.0 istatistik programı kullanılmıştır ve testin standart hatasını araştırmak için çevrimiçi istatistik hesaplayıcıdan yararlanılmıştır. Kapsam geçerliliği, testin ölçtüğü yetenekler ile ders kitabının müfredatında bulunan hedef ve kazanımlar ve testte kullanılan konuşma konuları ile ders kitabının konu kapsam ve içerikleri bir karşılaştırma tablosu kullanılarak araştırılmıştır. Yönetici, öğretmen ve öğrencilerin sınav hakkındaki görüşlerini öğrenmek için mülakat yapılmıştır. Faktör analizi yoluyla yapılan yapı geçerliliği analizi bulguları konuşma notunu oluşturan dokuz maddenin eigen değeri 6.42 olarak ve değişkenin %72.1'ini açıklayarak bir faktör altında gruplandığını göstermektedir.

Puanlayıcılar arası güvenilirlik analizi sonuçları 18 komisyonun 17'sinde puanlayıcılar arası korelasyonun pozitif ve güçlü olduğunu ortaya çıkarmıştır ($r_s \geq .69$, $p. \leq 0.03$). Puanlayıcı iç güvenilirlik analizleri 36 puanlayıcıdan 33'ünün puanlamasının yüksek güvenilirlikte olduğunu ortaya çıkarmıştır ($\alpha \geq .70$, 9 madde üzerinden). Ayrıca, testin standart hatası da $SE = .095$ olarak bulunmuştur. İçerik geçerliliği araştırmasında konuşma sınavında test edilen yeteneklerin ders kitabı müfredatındaki hedef ve kazanımların %70'i ile eşleştiği ve sınavda konuşulan konu başlıklarının ders kitabının konu kapsam ve içeriğinin %50'sini doğrudan test ettiği bulunmuştur. Mülakat sonuçları konuşma sınavının özgün, pratik olduğu ve öğrenci ve öğretmenler üzerinde olumlu bir etki oluşturduğu sonucunu göstermiştir. Sonuç olarak, geliştirilen bu sınav güvenilir, geçerli, otantik, pratiktir ve olumlu bir etki oluşturmaktadır.

Anahtar kelimeler: Geçerlilik, güvenilirlik, konuşma görev örnekleri, puanlama ölçeği, puanlayıcı eğitimi, yabancı dil olarak İngilizce konuşma değerlendirilmesi.

Table of Content

Certification	
Acknowledgement.....	i
Abstract.....	ii
Özet.....	iv
Table of Content.....	vi
List of Tables.....	ix
List of Figures.....	x
Chapter I: Introduction	1
The Context of the Study.....	3
Statement of the Problem.....	3
Subject of the Thesis.....	4
Purpose of the Thesis.....	4
The Significance of the Study.....	4
Concepts and Terms.....	5
Chapter II: Literature Review	7
Reliability in Testing.....	7
Validity in Testing.....	10
Testing Specifications.....	14
Speaking Tasks.....	18
Assessment Rubric.....	28
Research Studies on Speaking Exams.....	35
Chapter III: Methodology	45
Participants.....	46
Data Collection Instruments.....	47
Data Analysis.....	48
Development Process of Speaking Test.....	48
The construct of the speaking test.....	48

Selection of the test tasks.....	49
Assessment criteria.....	50
Procedures to conduct speaking test.....	51
Rater training.....	53
Piloting of the speaking test.....	55
Chapter IV: Findings	57
RQ1: How Valid Is the Test.....	57
1a. Does the test have content validity.....	60
1b Does the test have construct validity	63
RQ2: How Reliable Is the Test	64
2a. Does the test have inter-rater reliability.....	60
2b. Does the test have intra-rater reliability.....	63
RQ3: What Are The Views of Administrators, Teachers and Students About The Test....	66
Chapter V: Discussion.....	73
Discussion for RQ 1a and 1b.....	73
Discussion for RQ 2a and 2b.....	76
Discussion for RQ 3.....	78
Chapter VI: Conclusion and Implications.....	81
Pedagogical Implications.....	82
Methodological Implications.....	83
Future Research.....	85
References.....	86
Appendices.....	93
Appendix A: Scoring Sheet For Raters.....	93
Appendix B: Scoring Sheet for Taking Average Score.....	94
Appendix C: Sample Question for Sustained Monologue Production Task.....	95
Appendix D: Sample Question for Pair Interaction Task.....	96
Appendix E: Assessment Rubric for Task 1 (Sustained Monologue Production)	97
Appendix F: Assessment Rubric for Task 2 (Pair Interaction)	98

Appendix G: Official Permission Form for Using Online Sources of Cambridge English Language Assessment.....100

Appendix H: Official Permission Form for Using Online Sources of The Center for Advanced Research on Language Acquisition (CARLA) of University of Minnesota.....101

Appendix I: Official Permission Form for Using Q:Skills for Success Listening and Speaking 3 Course book of Oxford University Press.....102



List of Tables

Table No.	Title	P
1.	Academic Background Profile of Examiners.....	46
2.	The Questions to be Asked to Participants in the Interview.....	47
3.	A Comparison Table of Abilities Both Tested in the Exam and Taught in the Course..	57
4.	Comparison of Topics Tested in the Exam and Discussed in Classroom Sessions.....	59
5.	Mean Values, Communalities and Factor Loadings Based on PCA for 9 Sub-skills of Speaking Test.....	62
6.	Spearman Rank Order Correlation Analysis Between the 1 st Rater and 2 nd Rater.....	63
7.	Intra-rater Reliability of the 1 st and the 2 nd Rater.....	64

List of Figures

Figure No.	Title	P
1.	A task-specific rubric for a presentational writing task: “visiting monuments in Paris”-intermediate level.....	31
2.	An analytic scale of assessing writing performance scale in Cambridge PET examination.....	32
3.	An example of holistic scale of Cambridge PET speaking assessment rubric.....	33
4.	Primary-trait rubric.....	34
5.	Multi-trait rubric for interpersonal communication activity in target language.....	35
6.	Scree plot figure of eigen value of nine-item.....	62

Chapter I: Introduction

With the increase of interaction on global scale, especially in the last decade, the demand for learning a foreign language has erupted. Many institutions, now, are looking for employing agents who know foreign language(s). This search has caused to inclusion of foreign language education in education systems. Most of the universities have been providing preparation classes in which only a foreign language is taught.

English as a foreign language is the foremost foreign language taught in universities in Turkey. The teaching of foreign language is based on four skills which are namely reading, writing, listening and speaking. According to Ur (2000), for effective communication, among the four language skills, speaking is the forthcoming skill. Besides, Bailey and Savage (1994) emphasize that speaking is considered to be the basic skill. Moreover, Lazaraton (2001) regards that knowing a language has the similar meaning with to be able to speak it. Bearing the emphasis on the importance of speaking ability in mind, it could be deduced that speaking skill is very important and deserves a great quantity of attention in foreign language education.

Along with teaching of speaking skill, assessment of speaking is also crucial. Without conducting an assessment, it would be obscure to judge how much of the target aims and outcomes pertaining to speaking ability have been acquired by learners. Hughes (1990) indicates that it is difficult to measure oral ability accurately. Wei (2011) explains that form of language testing has to face the validity and reliability issues. In addition to that, O' Malley and Pierce (1996) indicates that the three challenges that are to be handled when assessing speaking are time setting, assessment tasks and evaluation criteria. It might be deduced that, selecting tasks that would elicit the speaking skills to be observed by examiners, applying criteria during the assessment by taking many dimensions of speaking skill into account

simultaneously in the restricted time by sustaining validity and reliability make the assessment of speaking skill very difficult.

Tasks are designs of many different contexts for classroom activities . Through tasks, test takers demonstrate their abilities under the given conditions. Therefore, tasks should be representative of the speaking abilities to be tested. As a conclusion, tasks are regarded to be important in testing speaking abilities.

On the other hand, evaluation criteria should be relevant to tasks and the abilities to be tested. Otherwise, since the performance of test taker and the assessment criteria would be two completely different things, scoring of the performance could be irrelevant of that test taker's real speaking ability. In addition, those criteria should be defined clearly and exactly enough for the examiners so that the same meaning could be understood and that could be applied precisely and accordingly by all examiners. Thus, a fair judgment could be provided for all test takers though raters change from commission to commission. As a conclusion, evaluation criteria could be considered to be significant in terms of scoring performances consistently.

Gong (2010) expresses that the way a speaking test is executed affects the validity and reliability of it. In other words, it could be inferred from the statement of the author, unless examiners apply the same procedures in terms of conducting speaking examination; for instance, time allotment, instructions, sequence of tasks etc., validity and reliability of the exam could be affected negatively. In addition to that external conditions such as heat, sound, comfort of desks, and quality of materials have impact on a valid and reliable speaking test.

As a conclusion, conducting a valid and reliable speaking test could be thought to be based on very delicate conditions. Definition of construct and implying the blueprint

flawlessly might have great significance. Therefore, there are many critical issues the needs to be taken care of when developing a speaking test.

The Context Of The Study

In the institution, during academic education students have 30 hours of English lesson in preparation class. English as a foreign language is taught based on four language skills namely reading, writing, listening and speaking in addition to main course lesson in which students are taught grammatical structures through communicative activities.

For assessment process, students take four mid-term exams and several quizzes. At the end of the education year, students have English Proficiency Exam. Mid-term examinations are comprised of four language skills, and the level of the examination is decided in the basis of proficiency level of education that students receive at that time. Quizzes are tests that cover unit subjects, aims and outcomes. Lastly, English proficiency exam is at CEFR B1 level and is based on four language skills.

Students' end of year score is calculated by taking 35% of the mean score of four mid-term scores, 15% of the mean score of all quizzes and 50% of the English proficiency exam score. As long as the accumulation of those scores is above 59, students pass preparation class.

Statement Of The Problem

When taking into account of the institution's necessity to hire agents who know English with a high degree of proficiency, the necessity of developing a valid and reliable speaking test in which speaking abilities of students could be assessed emerges out. During their careers, students will be in a large amount of different context in which English is medium of communication. Therefore, it is crucial to involve tasks in which test takers are

supposed to demonstrate many different speaking abilities and to create criteria for assessing the test takers' performance appropriately. Currently, there is not such an examination process.

Subject Of The Thesis

In this thesis study, how a valid and reliable speaking test is developed, which speaking abilities are to be tested, which type of tasks are included, which assessment criteria are used, the procedures for conducting speaking test are going to be elaborated.

Purpose Of The Thesis

In this study, a valid and reliable speaking test is aimed to be developed. In order to do that, abilities to be tested, tasks to be executed and criteria to be used for assessment will be decided. Next, the procedures to conduct the test and how to score performance appropriately will be presented to the examiners.

The Significance Of The Study

In this study, a comprehensive speaking test in terms of the abilities to be tested is aimed to be developed. In daily life, there are many contexts each of which has unique features. It holds great importance for students to graduate being capable of managing a conversation in those contexts. In speaking skill testing and assessment, it is expected from students to demonstrate those abilities through their performance. Therefore; this study is significant in terms of including a variety of speaking abilities to be tested.

Besides, in this study, it is important to define the steps to take in order to carry out the examination process successfully. In order to conduct the examination efficiently, deciding speaking tasks, preparing task related materials, instructions, planning speaking assessment commissions, objectives of raters, steps for undertaking the exam and other related documents

are going to be prepared and acknowledged to both raters and test takers. That point also adds uniqueness to the speaking examination.

Lastly, rating rubric is going to be prepared which is appropriate for the evaluating the relevant proficiency level abilities and raters will have a workshop session for rater training. In this respect, relevant materials will be prepared for that training. This section of the study is also very important in terms of sustaining consistency between raters, scoring validly and reliably and thus framing a rater-training model.

Concepts And Terms

The key terms that are considered essential in this study are presented as follows:

Assessor – “is a person who listens to a learner in an oral test and makes an evaluative judgment on what he/she hears (also examiner and tester).” (Underhill, 1987, p. 7).

Marker/ Rater/ Scorer – “is the judge or observer who observes a rating scale in the measurement of oral proficiency.” (Davies et al., 1999, p. 44).

Reliability – “is the consistency of evaluation of results.” (Gronlund & Linn, 1990, p. 48).

Rubric – “is a coherent set of criteria for students’ work that includes descriptions of levels of performance quality on the criteria.” (Brookheart, 2013, p. 4).

Speaking Test /Exam- is procedure in which a learner is assessed based on what he/she says in his/her speech.

Task – “are activities that people do, and in language-learning contexts tasks are usually defined in terms of language use.” (Luoma, 2007, p. 30).

Testee / Examinee/Candidate – is other different definitions for a test taker.

Validity – “deals with whether a test measures what it is supposed to” (Underhill, 1987, p. 9).



Chapter II: Literature Review

Assessment is a valuable part of the education/instruction process. Through assessment, learners are evaluated on the basis of how much they have achieved/acquired the proposed outcomes of the instruction. In this respect, tests are the instruments of assessing learners' performances. When learners take/undergo a test, their abilities/knowledge are demonstrated and become observable and measurable. Therefore, the development of a test holds great importance. There are basic considerations when a test is developed. The two indispensable criteria of them are validity and reliability issues. In order to carry out a satisfactory assessment by which it is meant the interpretations of the result of the test reveal what is intended to be measured, the two criteria ought to be sustained. Without a valid test, it becomes more than obscure to justify whether the results of the test actually demonstrate what the test taker is really capable of. Moreover, without a reliable test, the results of the test might not be dependable as it is unobvious if the results occurred by chance. As a conclusion, in order to carry out a dependable assessment, a validity and reliability issues have tremendous significance.

Reliability In Testing

Reliability is one of the most essential issues in testing and assessment which has several definitions. Genesee and Upshur (1996) state that the consistency of test results for the same individuals refers to reliability. The term is elaborated by them that a test rendering the same results for a given individual on different occasions would be judged as reliable. They also assert that if a test is not reliable, it cannot be valid. In another definition, reliability is defined as consistency of scores (Brown & Hudson, 2002; Henning, 1987). Luoma (2004) explains that if provided that the scores of a test which is given on the same day are reliable, they will stay almost similar if the same test is taken by the same people again. H. D. Brown (2004) indicates succinctly that a reliable test is consistent. Bruin (2010) states that a test is

thought to be reliable when a great many researchers can use it and they have the exact same results under same conditions. Based on the arguments stated above, the reliability could be defined results of a test occur out of chance and would repeat itself over time under the same conditions.

Reliability of a test is comprised of different dimensions. According to Genesee and Upshur (1996), types of reliability could be divided into test-retest reliability, parallel or alternate forms reliability, internal consistency and scorer reliability. Genesee and Upshur explain that the function of test-retest reliability is that the test would yield the same result on different times given to the same individual. By parallel or alternate forms reliability, they mean that similar results would occur for the same person with different but equivalent forms of the test. If the items that constitute a test correlate highly with each other, as they state, it demonstrates that the items are measuring the same skills and therefore test scores are consistent and it has internal reliability. Finally, scorer reliability is the consistency of the scores given by more than one scorer on the same test for the same test taker. As a conclusion, reliability of a test could be proven through different perspectives.

Sources of unreliability could stem from different reasons such as the test itself, testing conditions, test takers or raters (Brown, 2004; Genesee & Upshur, 1996). Genesee and Upshur (1996) state that the instructions of the test could cause confusion unless they are clear to be understood. In addition, the quality of printing or photocopy affects the legibility of the instructions or the questions as well. Moreover, H.D. Brown (2004) indicates that the atmosphere of the testing environment such as temperature, noise, comfort of desk and chairs, lighting conditions beside to emotional or physical condition of the test taker could affect the performance at the time of examination. He adds that in a testing occasion where there is more than one rater which is usually seen in speaking examinations, the issue of inter-rater reliability emerges. When scorers give inconsistent scoring to the same output because of the

possible reasons which are ignoring the scoring criteria, inexperience, inattention or even preconceived biases, inter-rater reliability, which according to Luoma (2004) means different raters rate performances similarly, weakens. H.D. Brown (2004) also adds that because of unclear scoring criteria, fatigue and bias, intra-rater reliability which, as he states, means that raters give consistent scoring to the same test even after a period of time under the same conditions could diminish. Regarding the arguments indicated, reliability is affected by several conditions which could cause a decrease in it.

There are several possible precautions to eliminate the above mentioned unreliability sources. The problems stemming from testing environment could be removed by setting comfortable desks and chair, providing noise insulated environment with proper lighting. Furthermore, announcing the time of examination before enough would enable test takers to be well ready for the exam. For the unreliability caused by raters, Luoma, (2004) indicates that the scoring instrument has a massive role in terms of re-establishing reliability and J. D. Brown (2005) adds that the analytical scoring instrument which is carefully specified can increase rater reliability. Another way of increasing rater reliability is through rater training programme. That could last for several days during which prospective raters are selected and go through a qualification procedure which includes independent rating of some taped-recorded performances, so that rating scores of them are consistent with ratings given by other qualified raters in the system. Luoma (2004, p. 177) explains the rater training session as:

Rating training sessions often begin with an introduction to the test and the criteria. Different levels on the scale are then illustrated, usually through taped performances that have been rated by experienced raters before the training. After this, the participants practice rating by viewing more taped performances. They report their scores aloud and discuss the reasons for the consensus score and any other scores that some of them might have given. (Luoma, 2004, p. 177)

As a result, there are possible ways to exterminate the sources of unreliability and thus contributing an increase in reliability.

There are different ways of judging whether a test or rating is reliable or not. Standard error of measurement is an indicator of the reliability of a test (Genesee & Upshur, 1996; Luoma, 2004). Genesee and Upshur (1996) elaborate that test scores might include an error component; thus, standard error of measurement is an indicator of how large the error component could be. Luoma (2004) states that as long as a test has high reliability, the standard error of measurement becomes smaller. Correlation calculation is another way of expressing reliability. Luoma (2004) indicates that “both Spearman rank-order and Pearson product-moment can be used to calculate both intra-rater reliability and inter-rater reliability.” (p.182). According to Butler (1985) and Cronbach (1990), values between .8 and .9 are usually accepted good values while .5 or .6 are considered worryingly weak in terms of inter-rater reliability. To summarize, there are different indicators of reliability which are standard error of measurement, Spearman rank-order and Pearson product-moment calculations.

Validity In Testing

Discussions on the validity of a test have been an ongoing issue over years. There are many definitions of the term validity. Hughes (1989) defines it as finding out if a test measures correctly what it is aimed for. Another definition by Henning (1989) is it is the relatedness of a given test or its sub-tests as a measure of what it is intended to measure. Almost as the same, Genesee and Upshur (1996) define the term as validity is how much a test can actually is able to measure what it is purposed to measure. Fulcher and Davidson (2007) deduce that validity is twofold one of which is there is an intention to measure something ‘real’ and the other is that whether a test actually does measure what is intended. As a conclusion, based on the definitions above, it could be inferred that a test is valid when it

actually measures appropriately what its aims are. To describe it in a more detailed way, in order to assure validity, a test developer needs to include questions or tasks that would elicit output from examinees that are aimed to be observed by examiners.

There are some dimensions of validity. Cronbach and Meehl (1955) describe these as; criterion-oriented validity comprised of predictive validity and concurrent validity, content validity and construct validity. From this point of view it could be deduced that validity is a multidimensional issue. Each type of validity contributes the validity of a test as a whole. In order to count a test as completely valid, all of the validity types should be sustained. These validity types obviously consolidate each other although each of them is not necessarily a prerequisite for the other.

Criterion oriented validity, as Fulcher and Davidson (2007) state, is that tester has interest in how much a particular test and a criterion are related to each other so that predictions could be made based on that relationship. H. D. Brown (2005) states that the scores on the criterion validated test should correlate highly with another, well relied measure of the same construct. Genesee and Upshur (1996) explain that criterion relatedness is shown by correlations between test score and criterion measures. It could be deduced that the new developed test needs to match to some extent with a previously valid test in order to obtain criterion validity. For example, if the results of the new designed test reveal the same results with previously valid test then the two tests measure the same criteria, or this new designed test could shed light on future capabilities of the test taker in terms of to what extent the test taker will be able to accomplish in the future tasks.

Predictive validity could be explained by an example. Assume that an institution would like to learn its effectiveness of a proficiency test of English given at the end of prep-class in future English-medium courses. As long as the students' achievement in the test matches with the performance shown in English-medium courses; then, it could be concluded

that the predictive validity of the proficiency test is high. Predictive validity, as Fulcher and Davidson (2007) state, is the term used when the results of the test are used to guess some future potential ability or success of a test taker. H. D. Brown (2005) also adds that the assessment criterion in such cases is to assess a test taker's probable future success. To estimate academic success at university of a test taker by judging his or her university entrance exam score, for instance, would be an example of predictive validity. Besides, speculating on the success of a test taker in lessons in which a foreign language is medium of education by looking at his or her foreign language exemption examination score could count for predictive validity. As a conclusion, predictive validity, could serve as an indicator of future potential of a test taker. In other words, based on the score of a test taker from a test, if a judgment can be made about possible success or failure in a situation about the test taker and it turns out to be true, then it could be concluded that the predictive validity of that test is high.

Concurrent validity could be explained with an example. Since it is difficult to arrange and administer a live interview of speaking test in terms of planning time, hiring trained interviewers, a semi-direct speaking test is conducted in which test takers' spoken performance is recorded to tapes to be later sent to trained raters for being scored by them. To see the effectiveness of semi-direct speaking test, scores on them can be correlated with scores on live interview to see criterion relatedness of the new test. This example is a study of concurrent validity. Fulcher and Davidson (2007) indicate that a new instrument is compared with a more established one which supposedly measures the same criteria when concurrent validity is investigated. They elaborate the term by adding that a well established test of the same ability needs to be given to the same test takers concurrently or after a few days passed the test which is supposed to be validated. Hughes (1990) only narrows down the time between the two tests by explaining that concurrent validity is established when the test and

the criterion are executed at the same time. In regard to the discussions above, it could also be estimated that it would be for the benefit of the designers of the new test to conduct the valid test in a short time so as to prevent environmental factors that could affect the performance of the test taker during the interval. However, the time between the two tests should not be too close for the test taker to have the second exam in case he or she might not lift the cognitive exhaustion. As a conclusion, it could be inferred that the results of the test to be validated and the valid test should be positively correlated in order to count the new developed test as concurrently valid.

During education, many different subject matters and skills are aimed to be gained. The way to find out to what extent those information and abilities are gained is to test the learner. Therefore, when developing a test, content of the examination ought to be so comprehensive as to include all aimed information and skills. As the name suggests, “content validity is about if the content of the test is represented sufficiently and is comprehensive enough for the test to be a valid measure of what is supposed to measure” (Henning, 1987, p. 94). In so doing, the validity of the new developed test would be increased as the result of the test would reveal more point of views about the performance of the test taker. Genesee and Upshur (1996) indicate that then a panel of experts in that field judges the relevance of the test content to the most current conceptions of that ability and add that content relevance is assessed logically, there is no statistical way in which it can be determined. As a conclusion, regarding the discussions above, it could be deduced that it is important to include as much of the course content as possible to sustain content validity. Thus, that test could be more comprehensive because of including more abilities to measure.

In a definition made by Cronbach and Meehl (1955), it is stated that if a construct is to be scientific, that has to be found in a ‘nomological network’, comprised of laws, statistical or deterministic. They elaborate the term that, these laws tie apparent parts to each other which is

to say the similar things or constructs are heaped up. Construct validity could be the first and the foremost issue when checking the validity issue; because, if a test measures what it intends to measure it could be decided that the test has the construct validity. The idea is also supported by Mislevy (2007) that, scientifically looking, construct validity is regarded as the complete of validity and forms the other types of validity. In addition to Mislevy, it is stated that construct validity has to be included whenever there is a measurement of certain criteria since the most practicable type of validity to evaluate measurements is construct validity (Andrews, 1984; Creswell, 2005; Mahoney, 2008; Messick, 1981, 1989; Popham, 2003; Embreston & Gorin, 2001; Gay & Airasian, 2003; McMillan & Schumacher, 2006). A definition made by Ebel and Frisbie (1991) states that construct validation is a process through which evidence is gathered to support the assertion that the test really measures what the developers of the construct aim to measure. Lastly, Messick (1989) explains that the test must be appropriate, related, and utilized accurately, with the crucial point being the integration of evidence that creates assumptions of the performance samples from the assessment results and he elaborates that these inferences must be clear enough to understand, dependable and help the aims of the assessment in order construct validity to achieve its purpose. As a conclusion, construct validity is concerned with the efficiency of a test to assess the knowledge about the relevant subject matter. The whole criteria or abilities aimed to be assessed indicate construct. Therefore it could be stated based on the discussions above, as long as a test is able to measure the construct, that test has construct validity.

Testing Specifications

It is very beneficial to write down the test specifications before developing a new test. Test specifications are defined as the 'blueprint' for a test, which are used by test and item writers to produce equivalent forms of the same test (Alderson, Clapham & Wall, 1995; Bachman & Palmer, 1996; Davidson, 2012). Furthermore, Fulcher (2010) indicates that it is a

single document that includes the purpose, construct, item types and the number of the item types of the test. Besides, Luoma (2007) explains that test specification consist of what kind of abilities will be focused on, how the outcome will be assessed and what kind of rating criteria will be utilized. In another statement, Fulcher (2003) emphasizes test specifications include test construct, a description of the tasks which are projected to be in the test, what kind of answers test takers are supposed to give, and an explanation of how the performances will be scored. Similarly with Fulcher, Luoma (2007) explains that test specifications contain construct definition to be assessed, and explanation of the test tasks and assessment criteria to provide guidance for developing similar tasks and for the fair ratings. The specifications underlie the rationale, the reasons for focusing on certain constructs in assessment, and how the tasks and criteria make them certain to operationalise. As a result, upon considering the elaborations stated above it could be deduced that test specifications are the attributes of a test which reflect the principles upon which the test is constructed to conduct examination and from which other tests that measure the same abilities, according to the same criteria and through same procedure could be produced.

Besides knowing what test specs is and how much it is useful in terms of developing a test, knowing how to write test specifications is also crucial. Fulcher (2003) regards that the process is difficult to describe since it is most of the time messy and the activities supposed to be included inside the test are heavily dependent on the testing context and he carries on by asserting that while it is not important to include all features in every test specification, many of them will be necessary across a wide range of context in which speaking in a foreign language is tested. According to Fulcher and Davidson (2007), early conceptions of test specification demonstrates normative views of test development which were prevalent back in time and the aim of those specs was to generate tasks which vary in terms of difficulty. It is also added that, two common elements exist in; sample test tasks and “guiding language”

about how to produce such samples. Bachman and Palmer (1996) recommend separate contents into two; the design statement including grounding definitions for the test, and the blue-print of the test involving the test specifications along with each of the task specifications which define the purpose, construct, setting, timing, instructions and linguistic features. Almost similarly with Bachman and Palmer, it is declared by Fulcher (2003) that specifications should be at two levels; one of which is the general specification the test containing procedures to carry out the test, parts of the test and their timing and the other is specifications which are more detailed for the tasks in the test. Lastly, Luoma (2007) states that the contents of her own model of modular specification are the same as in other models of specifications and adds that the major point is that specifications establish a record of background principles of the test. In summary, regarding the indications above, it could be inferred that while writing test specifications, it is significant to include general attributions pertaining to the test such as how to carry out tasks, and test task specifications including linguistic features.

There are many different models of test specifications. These specifications might demonstrate pathways for test developers. One of the earliest models presented by Popham (1978) includes general description; a succinct statement of the behavior supposed to be tested, prompt attribute; a full description of what the test taker will face, response attribute, a full description of suggested student's answer, sample item; a sample item or task that demonstrate the specifications, specification supplement; a full explanation of any further information required to construct items for a given spec such as a list of words for vocabulary specification.

On the other hand, another model depicted by Mislevy, Almond and Lukas (2003) involves item/task specification; illustrates the prompts designed to elicit answers based on which inferences are to be made regarding the targeted abilities of the test takers, evidence

specification; gives information about the kind of expected response, test assembly specification displays the number and range of items in the test, presentation specification; how the items and materials related to the test are to be presented, delivery specification; depicts the administration of the test, time constraints etc.

In addition, Alderson, Clapham and Wall (1995, pp. 11–20, 38) created a wide range list of test specs which involves:

The test's purpose, description of the examinees, test level, definition of construct (theoretical framework for the test), description of suitable language course or textbook, number of sections/papers, time for each section/paper, weighting for each section/paper, target language situation, text-types, text length, language skills to be tested, language elements to be tested, test tasks, test methods, rubrics, criteria for marking, descriptions of typical performance at each level, description of what candidates at each level can do in the real world, sample papers, samples of students' performance on task. (Alderson et al., 1995, pp. 11–20, 38)

Lastly, Luoma (2007) introduces her modular approach as construct specification that defines the combination of abstract skills to be tested with concrete implementation of them through tasks and criteria, task specification that includes detailed definition of each of the tasks such as item types, instruction about them and information of the tasks along with the administration process, assessment specification that defines the rating criteria and how they are to be conducted during the rating process.

As a conclusion, more or less, all types of test specification mentioned above share similarities in the list of specs although differ in terms of categorization. The fact that they differ in categorization should be understood as the difference in the point of view the way each spec should be dealt. That is to say, each spec is important and should be taken into consideration; however, some specs could be grouped and dealt with some other specs in a

more efficient way. Therefore, no matter which type of model among the ones above is chosen when developing a test, the specs that are recommended to be considered are almost similar.

The necessity of writing specifications might be questioned; however, documenting the decisions and the rationales behind them could be reasonable for several reasons. Fulcher (2003) states that the documentation has a role in validity argument as the constructs, tasks and rating scale are linked through a record of decisions. Similar to the Fulcher, Luoma (2007) emphasizes that a coherent system whose parts fit together emerges out by writing specifications. She adds that, the theoretical underpinnings would become more concrete for developers which results in a better understanding of the reasons behind the decision taking. As a conclusion, it would be beneficial for recording the test specifications as it would be clear for test developers how construct, tasks and rating criteria are connected and the rationale behind the definition of construct and selection of the tasks and rating criteria would be obvious for both test developers and the ones who are supposed to conduct the test.

Speaking Tasks

The way in which speaking is going to happen is influenced by the environment that surrounds. Luoma (2007) emphasizes that language use differentiate by purpose and context. In this respect, it could be inferred that the talk is adjusted either in informal or formal way, directed through monologue or dialogue, transaction or interaction. In language classrooms, these varieties of setting are provided through speaking activities for the learners so that they could have opportunities to engage in and improve their speaking skills in many different dimensions. It could be concluded that, the context defines the ways of speaking is going to happen.

Luoma (2004) indicates that when selecting the tasks the most significant issue is the construct-related information that the results should give. As mentioned above, with construct validity, what is targeted to be assessed through testing is aimed to be obtained. In other words, the qualifications or skills that language learners are supposed to have are tested via the speaking tasks. How far those skills are tested or the results of the examinees reveal what examiners want to learn about the abilities of the examinees are connected to construct of the test. As a result, it is important to bear the construct in mind when designing speaking tasks.

There are some other steps to be taken during the preparation of the task design. Luoma (2004) explains that in addition to the tasks, assessment designers would need to create the instructions for both the examinees and the interlocutors, the task related materials such as role-play cards or pictures. It could be deduced that all of these components are required to be designed in order to carry out the speaking tasks effectively during the examination. She adds that tasks could be categorized in many different ways according to the number of the examinees at a time, particular language use the examiners want to assess, approach of the assessment, combination of skills. The task design could be shaped based on the categorizations above. As a result, it could be concluded that task related instructions and materials and number of examinees at a time would dictate how the assessment will be made.

According to the number of the examinees at a time; the speaking examination could be categorized into three as individual, pair and group tasks. Luoma (2004) states that interview format is one of the most common way of setting speaking tests during which examinees are assessed at a time. This type of testing examinees contains advantages and disadvantages. It is flexible that, as Luoma (2004) declares, the questions can be adapted to each examinee and the testers can control things happening dominantly during interaction. However, this excessive control is accepted as a weakness as well (Bachman, 1998; Lazaraton, 1992; Savignon, 1985; van Lier, 1989). It is implied by the authors that while the

interlocutor manipulates all segments of the interaction, the role of examinee is to comply and reply. It could be inferred that, the examinees might not know how to carry out, manipulate and direct a conversation. In conclusion, it could be stated that in interview task, the interlocutor might ask several questions to elicit output wanted to be observed.

Compared to individual interview, pair interview surpasses it in some features. Swain (2001) puts forward three arguments in favor of conducting pair interview. The first one is more different types of output is desired to be elicited than in the conventional interview and as a result enlarge and enrich the evidence obtained about the examinees' skills. Secondly, she continues, is about the relationship between testing and teaching, either in the sense of creating a positive washback resulting in the influence in classroom practices in order to encourage pair work much more in the classroom, or in the mood of rehearsing what has been happening in classroom up to that time during the test. Lastly, the third reason is economical; the amount of examiner time necessary for executing the tests is decreased because of interviewing pairs. Luoma (2004) states that the way it is conducted leaves more initiative to the examinees who interact with each other in the main part of the test than the examiner who observes the conversation. On the other hand, paired tasks bring some disadvantages with it. The concern is that all test takers might not have same amount of opportunity to demonstrate their best speaking performance (Iwashita, 1999; Weir, 1993). Because of the over dominance of one of the pairs, or roles of the pairs, imbalance might occur in terms of speaking time. However, when attributes of different pairs' effects on the scores is inspected, the results depict that the influence is small and even sometimes some other studies revealed contradictory results about which attributes inflate or shrink them (Berry, 1997; Iwashita, 1999; O'Sullivan, 2002). Luoma (2004) asserts that this might occur since there are a lot of variables that their interrelated effect on test takers is hard to foresee. Furthermore, since raters probably pay attention to the different features of interaction, this could influence their

ratings in a variety of ways. Another feature of paired task, as Luoma (2004) states, that the examiners mostly feel unsure about the amount of responsibility provided to the test takers who have not been trained in paired interview techniques and adds that and the speaking test should contain so clear task materials and instructions that it could enact the discussion and the examinees are able to understand what kind of performances will help them have better scores. Based on the discussions above it could be concluded that, although paired interview tasks might seem disadvantageous in terms of the possibility of one of the participant's being dominant and therefore not giving enough opportunity for the other, it is advantageous in terms of giving examiners enough opportunity to observe different types of output.

Similar to pair work, group interaction tasks are also mostly well accepted by learners (Fulcher, 1996; Shohamy, Reves and Bejenaro, 1986); however, Reves (1991) indicates that probably because of the administrative concerns related to handling the sizes of groups and the abundance of proficiency levels in them, they are not mostly implemented in formal tests of speaking. It could be driven that should the sizes are too large for one rater to score the performances of the examinees, it would be difficult to focus on them and thus misjudgments could arise. Besides, it could also be deduced that roles and responsibilities of the examinees are important as in pair work; because, they should understand clearly how to carry out the task and end it. Lastly, it could also be inferred that, in a case of unbalanced levels of examinees of a group, the one with lower level could fail to fulfill his/her role during the task and as a result affect the others negatively in terms of following the procedures and instructions of the task. As a conclusion, it could be stated that, group tasks might have the same advantages and disadvantages with pair work; however, the most important issue in creating a speaking task is to clarify the roles and objectives of group members to carry out the task.

Nunan (1989, as cited in Luoma, 2004, p.40) separates the speaking tasks into two in terms of language use that the examiners want to assess; “pedagogic or language focused and real-life or target tasks”. Pedagogic tasks, as defined by Nunan (1989), are developed especially for certain limited types of language use. Luoma (2004) elaborates the pedagogic tasks that there is even less than slight relationship between the activity and real-life instruction, although it is communicative and meaning focused she continues by exemplifying that it makes one of the learners instruct the other such as instructing to draw picture and instructing examinee with the examiner work together in inspecting if the instruction receiver is able to follow. Nunan (1989) defines real life tasks as tasks which simulate language use outside classroom. Luoma (2004) explains that real life tasks demonstrate the fundamentals of non-test specific language use in the assessment situation and this is generally carried out through simulation or role-play. She exemplifies that examinees are put in professional role in typical real-life tasks while the examiners act as clients, visitors etc. in which examiners might have interaction with the test takers in occupational contexts. McNamara (1996) pulls attention to delicate balance between the linguist’s perception of required ability of language use and the professional’s perception of relevant professional communication in view of designing tasks of real-life test. He categorizes performance testing into weak and strong. Strong performance testing simulates the events in which real-world use of language is displayed and also includes real-life assessment criteria for evaluating examination success, however in a weak performance test, evidence of displaying sufficient language ability would be adequate to earn a satisfactory rating. As a result, it could be concluded that while pedagogical/language use tasks aim to elicit language output to be assessed in the light of pedagogical purposes, real-life/target tasks aim to elicit language output related to real-life context specific language features.

The approach that is to be obtained for creating speaking test tasks could also be decided through intended use of the scores of the examinees. When the object of the assessment is to test the language ability in a broad sense it is called construct-based approach because, as Luoma (2004) states, the main focus is on the construct of language ability. She explains that with the construct-based tests, the construct is defined in the basis of course syllabuses, theoretical models, needs analysis, next the developers should provide content-, construct- and process-related samples that the language skills needed to be demonstrated by the test taker in test performance and stated by the raters when assigning scores would comply with that definition. On the other hand, Long and Norris (2000) remark that task-based language assessment prioritize the task itself as the principal object to be analyzed, inspiring the selection items, construction of test materials and the scoring of the performance related to the task. Ellis (2003) indicate that task-based approach is regarded as a course of discovering the correlation between the test performances, such as what the examinee is doing in the speaking exam, and the assessment criterion of performance, such as what the examinee must do in the context of real life. Another elaboration about how the task based approach should be adopted in language testing is made by Luoma (2004) that the developers should provide that the content of the tasks demonstrates the demands of the related task apart from the examination context, and that the rating should show that. Regarding the statements above, it could be deduced that a task based approach in assessment would require the use of simulations and strong performance testing should be conducted in the speaking examination so as to replicate real-world context in the test. On the other hand, in construct based approach, pedagogic tasks in relation with course syllabus could be included in order to assess the quality of language use. It could also be added that, however, real life tasks or simulations might also be utilized as long as weak performance testing is applied this is because the language ability that an examinee is expected to demonstrate would be enough to gain a good

score or pass the test at least as mentioned above. This view is also supported by Luoma (2004) in some way who states that the assessment criteria in many language test employ the idea of weak performance testing, possibly as both the examiners and the test takers bear in mind that a language test is conducted for assessing language use. As a result, it could be concluded that when adopting task based approach, the test results would yield the possible capability of the test taker in real life situations and therefore both tasks and scoring criteria would be shaped accordingly; however, when adopting a construct based approach the test results would yield interpretations about the test taker's speaking ability.

Speaking tasks can also be designed as integrated tasks or independent tasks. Plakans (2017) defines integrated assessment as combination of more than one skill in a test, such as reading/writing or reading/listening/speaking. In this type of assessment in speaking examinations, examinees are provided a reading or listening material upon which they are expected to develop the topic or reflect their opinion. Independent tasks require the exclusion of other language skills except speaking. Nakamura (2014) defines that independent tasks demand from examinees to rely on their personal background knowledge to undertake the task. There are different views about integrated and independent tasks. Read (1990) and Weir (1993) advocate for the usage of integrated tasks stating that test takers might lack sufficient information on which to build their argument. In addition, Wesche (1987) claims that validity would be increased through replicating real life communication tasks as in academic contexts. However, Luoma (2004) points out that the examinees did better in standalone tasks than in the integrated, probably because of the enhanced cognitive load in integrated task. As a conclusion, both integrated and independent tasks seem to have advantages and disadvantages; when choosing which one to include in speaking test, it would be a better idea to look construct definition of the exam.

In speaking examinations, speaking tasks are important in terms of requiring speaking skills that examiners would like to observe. As stated above, speaking is influenced by context resulting in different types of speeches. Occasions change in daily life so many times that it could become necessary to communicate in these different types of speeches. Therefore speaking tasks in the examination should reflect the varieties of daily life speaking in this sense. In a study conducted by Brown and Yule (1983), distinctions between four types of informational talk are illustrated as giving instructions, description, telling stories and opinion expressing. Furthermore, Bygate (1987) made even finer separations among types of speaking tasks as information based speech which are describing an event/picture etc., narrating a story/memory etc., giving instructions, comparing and contrasting two things and evaluative talk which are explaining/justifying ideas, predicting, and decision making. It could be inferred from Bygate's distinction that in factually oriented speaking tasks, examinees are expected to structure and produce their speech based on the instruction of the task. However, in evaluative tasks, the examinees also rationalize their thoughts while performing the task by making reasoning. Thus, the varieties of these tasks require the demonstration of different speaking abilities. Below, more elaboration is provided about these speaking types. As a conclusion, there are many contexts in real life and the classification of the types of speaking tasks could be made based on those contexts so that selection of tasks for a speaking test developer could be easier as that kind of categorizations could shed more light on what kind of speaking abilities would be elicited.

In description tasks, as Kayi (2006) defines, examinees are given most of the time a picture and they are asked to describe, give as many details as possible about the picture and if that is a pair work with examinee or interlocutor, for example, they can describe the pictures to each other and find the similarities and differences or place missing items etc. The task also could require the examinee to describe his home, place etc. As a conclusion, in

description tasks, giving detailed descriptions such as what is/are seen in the picture(s) and where they are exactly, the shapes/colors numbers of the objects, actions happening in the picture etc. in an organized way would be in favor of the examinee.

Narrative tasks, Fulcher (2003) states, are often used to test the ability to sequence events, and assess the test taker's ability to control time markers and past tense structures. He elaborates that examiners should ask the examinees to demonstrate their command of the attributes of narration such as setting the scene, introducing the characters, mentioning to those characters accurately and appropriately, figuring out the major events and talking about them in an order coherently. As a conclusion, it could be inferred that narrative tasks would require a test taker to show the control of the genre, and to structure the discourse.

According to Goh and Burns (2012), explaining and predicting tasks could either be monologue or interactive. They maintain that explaining and predicting tasks are very suitable for students studying in academic programs or in ESP programs and they require the student to explain the information often found in graph or table form and extrapolate from the information to explain or predict the meaning. They add that, to complete the task, students should explain the background, identify the components of the information or processes and organize them in a coherent order so that listeners understand the information. As a conclusion, it could be deduced that, examinee is supposed to analyze the situation and form a step based on which s/he is supposed to make elaborations in a discourse.

Decision making tasks involve the discussion of an issue from more than one point of view (Luoma, 2004.). She continues that test takers, then, come to the best decision and if the test taker is tested individually, s/he is supposed to articulate various points of view and then give a reason why one of them is chosen as the final decision. As a result, it could be inferred that mostly similar to explaining and predicting task, the test taker needs make a good reasoning and then finalize his or her speech in a discourse.

Role play tasks enable students take on a role rather than talking in a monologue fashion, and they therefore, supply scope for facilitating students to perform abilities of language use for the role. Fulcher (2003) states, simulations are similar to role plays. It could be deduced that role plays could align more closely with real life tasks that students need to carry out. Usually, students are provided with a situation within a specific context, some guidelines about what they should discuss. As a result, it could be concluded that role plays provide a situation in which test takers' interaction abilities are displayed in a wide variety of contexts and thus enabling different kinds of output.

In compare and contrast type of tasks, Stirling (2012) indicates that, examinees are required to contrast two ideas, pictures, define each idea, give an illustration of each, then compare and contrast them. In doing so, Luoma (2004) adds, examinees are expected to give in depth information and discuss both similar and different aspects. It could be deduced that this might require the use of forms for comparing and contrasting, conjunctions and complex structures and test takers might need to analyze both prompts and talk about them in a discourse.

In giving directions or instruction tasks, examinees are expected to get the message across and make sure that it has been received (Luoma, 2004). She adds, while it could be interactive in a dialogue mode, it could also be in a one way direction speech. It could be deduced that when it is in interactive form, direction instruction can repeat all the instructions at the end, or he just can simply ask questions to have the instruction giver make more exact instructions. As a conclusion, giving instruction task is a controlled task in which test taker needs to show that s/he comprehends the directions and produce the expected output.

In information gap activity, Kayi (2006), students are expected to work in pairs. One student partners will share their information through asking questions to find missing information. She explains further that Information gap activities provide opportunities to

solve a problem or collect information. As a conclusion, these activities are effective because partners could talk in the target language by constructing a dialogue.

In conclusion, speaking tasks should not be limited to the tasks explained above. They are the ones mostly used in speaking examinations or in classroom use. It could also be inferred that by employing some of the tasks above in a speaking test, an approximate judgment could be made about the one's speaking skills. Adding variety to the tasks included in a speaking test could increase validity in terms of the variety of the abilities tested; however, the increase in number could add the cognitive burden while dealing with the tasks for test takers and assessment for raters. Therefore, keeping a balance is an important issue while deciding the variety and number of the tasks when developing a speaking test.

Assessment Rubric

In the view of Fulcher and Marquez Reiter (2003), assessment scales endow operational definition for linguistic construct involving the aspects of speaking. Another definition by Davies et al. (1999) is that scoring rubric is a scale for describing language proficiency which includes a series of levels according to which the performance of test taker is assessed. In addition to the definitions above, Luoma (2007) explains that a rating scale which contains categories of numbers or verbal statuses along with the descriptors for stating what each number or verbal status stands for is an asset used for determining how well examinees can speak. In conclusion, the idea that assessment scales include levels which describe performance features from least successful to most successful, and are utilized for assigning a score for the speech given by the test taker could be drawn.

Scales could be found in different formats. There have been many scale forms designed for different purposes. Alderson (1991) implies that there are scales which have different uses such as user oriented scales which are used to convey information about typical output of a test taker at a given level. Fulcher (2004) indicates that it is an appropriate way to

note down the descriptors of level in view of what the learner is able to do in the foreign language, or in 'Can Do' statements. The other use of scale is, as Fulcher (2004) states, assessor-oriented scale designed to guide the process of rating, converging the quality of the performance expected. Lastly, constructor-oriented scales created to aid the constructor of test choose test tasks to be selected. According to Fulcher (2004) "they contain references to the types of task that are most likely to elicit the language sample required for the scores to be meaningful" (p. 89). As a result, it could be concluded that, the aim of the assessment might determine the selection of the scale type. If certain type of speech output is expected then it might be useful to have user oriented scale, if the quality of the speech output matters then it might be useful to have assessor oriented scale, if task specific output is expected then it might be useful to have constructor oriented scale.

Apart from the categorization by Alderson (1991), there are other categorizations for rubric or scale types. General (or generic) type of rubric, for example, is the one used across similar performances such as all writings, all oral presentations, all group interactions (Arter & McTighe, 2001). The idea is elaborated as using the same writing assessment scale for different kinds of writing performances no matter what the writing task is. Similarly, Brookheart (2013) declares that generic rubrics use criteria and performance descriptions that could be generalized across different tasks. She explains that all of the tasks have to reflect examples of the same learning outcome. Based on the definition given above, it could be regarded that general, or generic, rubric can be used for different tasks in terms of the same assessment criteria. An example for general rubric is given below by Brookheart (2013):

The 6+1 Trait Writing rubrics were developed in the 1980s by teachers working with the Northwest Regional Educational Laboratory, now Education Northwest (educationnorthwest.org). Identifying the six traits was a large part of that work. The six (plus one) traits are the following:

- Ideas
- Organization
- Voice
- Word Choice
- Sentence Fluency
- Conventions
- Presentation

(Brookheart , 2013, pp. 42-43)

Another type of rubric is task-specific rubric which as stated by Moskal (2000) is used for evaluating the performance of student on a single assessment event. She expands the term that it might be required to develop a different scoring rubric when tasks in a subject differentiates. Besides, Tedick (2002) expresses that task-specific rubric is used for considering the task foremost granting that the scale demonstrates a good correspondence with the task and test construct. Considering the definitions given above, it might be decided that if a task-specific rubric is to be used, the rubric should include assessment criteria uniquely for the achievement of the task; therefore that rubric can only be used for that task only. An example for the task-specific rubric is provided in Figure 1 below:

**Figure 2. Task-specific Rubric for a Presentational Writing Task:
"Visiting Monuments in Paris" - Intermediate Level**

	Strong Performance	Meets Expectations	Approaching Expectations
Use of past tenses (Domains: Functions, Language Control)	Past and imperfect tenses are used appropriately, and forms of all verbs are accurately spelled. Any error in choice of tense/spelling is minor and infrequent, and does not interfere with understanding.	Past and imperfect tenses are used appropriately most of the time. Verb forms are accurately spelled most of the time. Errors do not interfere with understanding.	Choice of past and imperfect tenses is inappropriate in several sentences and/or several verb forms are misspelled. Errors make understanding difficult at times.
Use of story form (Domain: Text type)	Story is well-organized with descriptions and details to add interest. A variety of sequencing words are used to connect the events of the story appropriately.	Story is well-organized with some descriptions and/or details to add interest. Sequencing words are used to connect the events of the story.	Story is a list of sentences loosely connected with some sequencing words.
Use of cultural knowledge (Domain: Cultural Awareness)	Incorporates extensive and correct knowledge of current and historical significance of all monuments pictured.	Incorporates correct knowledge of current and historical significance of all monuments pictured.	Incorporates a minimal amount of correct knowledge of current and/or historical significance of all monuments pictured.

Figure 1: Task-specific rubric for a presentational writing task: "Visiting monuments in Paris" - intermediate level (CARLA of University of Minnesota, 2019, para. 1).

One of the other types of rubric is analytic rubric. Thornbury (2005) explains that "analytic scoring is to give a separate score for each kind of aspect of the task" (p. 127). He adds that, provided that the factors in the rubric are well selected, the scoring would be fairer and more reliable. On the other hand, he continues, the rater could lose concentration while focusing on all of those aspects. Apart from Thornbury, Brookheart (2013) indicates that "analytic rubric describes the work on each criterion separately" (p. 6). She implies that using analytic scoring could be useful in terms of showing learners what kind of feedback should be given to improve the underachieving aspect of their work. However, Brookheart and Nitko (2008), and H. D. Brown (2004) explain that it is harder to achieve inter-rater reliability when using analytic rubric. Lastly, Tedick (2002) states that different weightings can be given for different aspects, thus that enables teachers to prioritize the aspect that they judge as more important. Upon considering all the statements about analytic rubric, it could be deduced that analytic scales include different aspects of a task separately as assessment criteria and have both advantages and disadvantages. An example of an analytic scale is given Figure 2 below:

B1	CONTENT	COMMUNICATIVE ACHIEVEMENT	ORGANISATION	LANGUAGE
5	All content is relevant to the task. Target reader is fully informed.	Uses the conventions of the communicative task to hold the target reader's attention and communicate straightforward ideas.	Text is generally well-organised and coherent, using a variety of linking words and cohesive devices.	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.
4	<i>Performance shares features of Bands 3 and 5.</i>			
3	Minor irrelevances and/or omissions may be present. Target reader is on the whole informed.	Uses the conventions of the communicative task in generally appropriate ways to communicate straightforward ideas.	Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control. While errors are noticeable, meaning can still be determined.
2	<i>Performance shares features of Bands 1 and 3.</i>			
1	Irrelevances and misinterpretation of task may be present. Target reader is minimally informed.	Produces text that communicates simple ideas in simple ways.	Text is connected using basic, high-frequency linking words.	Uses basic vocabulary reasonably appropriately. Uses simple grammatical forms with some degree of control. Errors may impede meaning at times.
0	Content is totally irrelevant. Target reader is not informed.	<i>Performance below Band 1.</i>		

Figure 2: An analytic scale of assessing writing performance scale in Cambridge PET examination (UCLES, 2014, p.2).

Apart from analytic rubric, there is also holistic rubric. Fulcher (2004) defines it as a rubric by which a single score is assigned based on impressions or according to the descriptor it has. He states further that this single score entails all features of the performance or output, demonstrating the overall quality. Besides Fulcher, Thornbury (2005) indicates that “holistic scoring is to give a single score based on an overall impression” (p. 127). Moreover, Brookheart (2013) defines that “holistic rubrics use all the criteria at the same time to describe the output thus enabling an overall judgment about the quality of the output” (p. 6). There are some advantages for using holistic rubric. In addition to sustaining inter-rater reliability (Brookheart & Nitko, 2008; H. D. Brown, 2004), holistic rubric provides advantage of quicker scoring (Thornbury, 2005). On the other hand, holistic rubric lacks specific details; therefore, test takers are unable to see their weaknesses exactly (Tedick, 2002). In conclusion,

it could be inferred from the statements above, holistic rubrics enables raters give a single score for the whole performance based on overall impressions including both advantages and disadvantages. An example of a holistic scale is given Figure 3 below:

B1	Global achievement
5	Handles communication on familiar topics, despite some hesitation. Organises extended discourse but occasionally produces utterances that lack coherence, and some inaccuracies and inappropriate usage occur.
4	Performance shares features of Bands 3 and 5.
3	Handles communication in everyday situations, despite hesitation. Constructs longer utterances but is not able to use complex language except in well-rehearsed utterances.
2	Performance shares features of Bands 1 and 3.
1	Conveys basic meaning in very familiar everyday situations. Produces utterances which tend to be very short - words or phrases - with frequent hesitation and pauses.
0	Performance below Band 1.

Figure 3: An example of holistic scale of Cambridge PET speaking assessment rubric Retrieved from: (UCLES, 2016, p.61).

As indicated above, rubrics can also be designed for assessing task specific features as well and primary-trait rubric is one type of task-specific rubric (Lloyd-Jones, 1977). He expresses that it was created to assess the main functions of language or specific trait which is produced by the test taker through the task. Similarly, Tedick (2002) remarks that primary-trait scoring requires limiting criteria on a task to one underlying dimension when assessing performance. He continues that it is a faster way to score as well. However, Applebee (2000) notes that in order to assure the raters solely focusing on whether the test taker accomplishes the purpose of the task, raters are instructed to ignore errors and concentrate on overall rhetorical effectiveness. As a conclusion, it could be understood that when using a primary-

trait rubric, it is important to taking only one criterion specific for the task into consideration.

An example of a primary-trait rubric is given Figure 4 below:

<i>Primary Trait: Persuading an audience</i>	
0	— Fails to persuade the audience.
1	— Attempts to persuade but does not provide sufficient support.
2	— Presents a somewhat persuasive argument but without consistent development and support.
3	— Develops a persuasive argument that is well developed and supported.

Figure 4: Primary-trait rubric (Tedick, 2002, p.36).

Hamp-Lyons (1991) introduces another type of task specific rubric as multi-trait rubric. Tedick (2002) defines that multi-trait rubrics permit the raters to assess the performance in the view of more than one dimension of task-specific features. He elaborates that, multi-trait task specific rubrics seem similar to analytic rubrics in terms of including several dimensions to score; however, it differs from analytic rubrics in terms of the task specific criteria it includes. It might also be developed as holistically since holistic rubrics require raters to assign a score based on overall performance while holistic rubrics include several criteria for the assessment. As a conclusion, it might be inferred that, if the task is to be evaluated from several task-specific features, multiple-trait rubrics could be utilized. An example of a multiple-trait rubric is given Figure 5 below:

Primary Activity Standard: Communication Standard 1.1 (Interpersonal Communication)

Students engage in conversations, provide and obtain information, express feelings and emotions, and exchange opinions.

	Excellent	Average	Needs Work
Time on Task	The group forms immediately to work on activity until the teacher indicates otherwise; if group finishes early, members discuss topics related to TL. 10 9	The group forms fairly soon to work mostly on activity until the teacher indicates otherwise; if group finishes early, members are either silent or discuss topics not related to TL. 8 7 6	The group takes a long time to form; they do not work on activity (unless the teacher walks by); if group finishes early, members discuss topics not related to TL. 5 4 3 2 1 0
Participation	All group members participate equally throughout the entire activity. 5	All group members but one participate equally throughout the activity. 4 3	More than one group member does not participate equally throughout the activity. 2 1 0
Group Cooperation	All members cooperate to help each other learn; if anyone has been absent, the group helps him/her; no one acts "superior." 10 9	Most members cooperate to help each other learn; if anyone has been absent, the group sometimes helps him/her; no one acts "superior." 8 7 6	Members do not cooperate to help each other learn; if anyone has been absent, the group does not help; some members act "superior." 5 4 3 2 1 0
Use of TL	Members use as much TL as possible (also to greet and say farewells). 5	Members use some TL during activity (also to greet and say farewells). 4 3	Members rarely use TL during activity (neither do they greet nor say farewells). 2 1 0

Figure 5: Multi-trait rubric for interpersonal communication activity in target language (Petersen, 1999, as quoted in CARLA of University of Minnesota, 2019, para. 1).

Research Studies On Speaking Exams

In a study conducted by Zhou (2016), it was aimed to investigate construct validity of the speaking examination which tests the test takers' communicative proficiency in Beijing International Studies University. TEP (Oral) Level B speaking examination aims to elicit language samples such as indicating factual information, explaining personal ideas and reasoning along with demonstrating competence with being sufficiently fluent on familiar topics. The exam is comprised of three parts; first of which includes questions about test taker's personal background, the second part includes text retelling and question and answer session related to that. In the final part of the examination, test takers are evaluated on their co-performance through oral discussion. 40 percent of a test-taker's speaking performance evaluation is defined based on holistic scoring by interlocutor, and 60 percent of it is defined through analytic scoring by assessor involving sub-skills of interaction, organization of content, pronunciation, grammar and vocabulary. The study included 36 trained raters and

254 test-takers. Reliability analysis of TEP (Oral) B revealed a Cronbach's alpha score of ($\alpha = .93$) over 5 items. This result shows that the scores of test takers in TEP (Oral) B are highly reliable. Besides, in order to check to what extent inter-rater reliability was achieved, Spearman's rank correlation coefficient is conducted. The correlation value of the scores of seven pairs are $r < .65$, and of the eleven pairs are $r > .65$ and all have $p < .01$ values. Upon considering the statement of Butler (1985) and Bachman (1990) which is .5 or .6 range are considered weak correlation; it could be inferred that there are reasonable correlations between the scores of eleven pairs out of eighteen. According to Xiaoqing (2003), factor analysis could be used in finding construct validity and the researcher utilized SPSS.22 in doing so. The results indicated that the KMO coefficient was .886 which was an acceptable value to administer FA as "a value of .60 or higher is accepted to be sufficient" (Razı, 2012, p. 174). According to the communalities results, there appears only one factor and all factor items have factor loads of $> .450$; therefore, regarding the statement of Büyüköztürk (2007) there is no need for extraction of the any item. Finally, the initial eigen values indicated that the first factor explained 77.3% of the variance, which is solely enough for being a construct since Razı (2012) considers 60% of cumulatively explained variance as highly sufficient. Thus, it could be concluded that the TEP (Oral) Level B exam has construct validity. In addition to that, the researcher carried out multi facet rasch models analysis since MFRM could reveal more information about interactions between a specific rater with a task or a test-taker and contribute more to investigation of construct validity (Grabowski, 2009; Lynch & McNamara, 1998; McNamara & Roever, 2006; Rasch, 1960). MFRM has also indicated that although raters used the scale appropriately, upon taking the test-takers' abilities into consideration the tasks at Level B could be regarded as relatively easy for most of them, the tasks of text retellings and topic discussions ought to be restructured thus the tasks can accurately show the language abilities of test takers. As a conclusion, the speaking

examination included question and answer session, retelling a text and question and answer session based on that and lastly an oral discussion with another test taker. Assessment criteria included an analytic scoring on interaction, organization of content, pronunciation, grammar and vocabulary. Cronbach alpha reliability analysis of the exam result is high ($\alpha \geq .90$, for each of the sub skill evaluated in the exam), inter-rater reliability analysis demonstrates that 11 pairs out of 18 is satisfactorily and positively correlated ($r_s \geq .67$) based on Butler's (1985) and Cronbach's (1990) statement above. Construct validity of the exam is positive after Factor Analysis (FA) which showed that the first factor explained 77.27% with eigen value of 3.86. and four items (communicative effect, text retelling, topic discussion, vocab & gram, pron. & into.) have a factor load of > 0.67 in the first and the only component. However, based on Multi facet Rasch analysis, the validity of the exam is low since tasks were too easy for the test takers to reflect their real potential. In the study, there are no signs of content validity and criterion related validity.

In another study by Zhihong, Zhenxiao and Leijuan (2016), it was aimed to find the how beneficial it was to conduct English speaking test with computer assistance in the scheme of Communicative Language Testing (CLT). In the study, the research questions of how much valid and reliable the computer assisted Spoken Language Speaking Test (SET) from the CLT model perspective and how much effective such a computer-assisted SET in reflecting EFL learners' real language use abilities are tried to be found out. The number of participants was 34 juniors of non-English major students. The exam included three tasks. The first one is group discussion in which students one by one lead discussion of the topics by ensuring the participation of every one of the students in discussions by including related communicative strategies taught during the course. The next one is a conversation in a pair work lasting for five minutes. The last task is personal speech about the topic talked previously. The assessment criteria were a blend of IELTS Speaking Band Descriptions and national College

English Test-Spoken English Test (CET-SET); however it is not supplied in the study paper. To obtain content validity, test specifications and the sub-tests were compared to show to what extent they match consistently. It was found by the researchers that first task allows each participant in the group to perform their communicative competence equally. Additionally, in second task, sociolinguistic and illocutionary competence contained by pragmatic competence of the test takers are able to be evaluated since the task required them to exchange ideas and facts within the target language by relating the topics to their life experiences thus involving socio-cultural factors inevitably. Lastly, in the third part of the test, organizational competence including linguistic and textual competence of the test takers is assessed. As a result, it could be understood that, the exam has content validity. In order to investigate the construct validity, Pearson correlation coefficient calculations were conducted to depict the correlations among sub-tests, and between sub-tests and the test as a whole. The correlation of group discussion, pair work and personal statement between the total test score were positively correlated as $r_p(34) = .73, p < .01$, $r_p(34) = .81, p < .01$, $r_p(34) = .72, p < .01$ respectively. Therefore it could be concluded that there is a construct validity of the test as there is significant positive high correlation between the sub-tests and the total scores as expressed by Fulcher (2003). In order to find reliability of the scores, Cronbach Alpha coefficient is calculated. The result is Cronbach's alpha score of ($\alpha = .84$) over 34 scores which means that the test scores have high reliability. As a conclusion, the study included a group discussion, pair discussion and personal speech as speaking tasks. The assessment criteria were a blend of IELTS and a national scale called CET-SET. In order to find content validity of the exam, tasks and test specification were compared and as a result consistency between them is assured. In order to check reliability, Cronbach alpha analysis was conducted and a high reliability was found. Lastly, in order to find construct validity Pearson Correlation

Coefficient analysis was carried out between the results of test tasks and overall test score. The results were all high significant positive correlations which indicated construct validity.

In a master thesis study conducted by Sak K1ymazaslan (2008), validity and reliability of the speaking test was aimed to be investigated which is conducted at a Turkish University. The speaking exam consisted of two tasks one of which is picture description and the other is topic explanation task. In the study, there were 70 test taker participants and six raters. The data were collected through interviews, questionnaires, exam results of speaking, TOEFL and departmental speaking exam. In order to find face validity, the results of the questionnaire were used and interviews were conducted to examine the content validity of the exam. The questionnaire results revealed that the exam had face validity at a satisfactory level. Furthermore, the interview results indicated that the exam had content validity at a relatively high degree level. In order to find predictive validity, Pearson correlation coefficient was conducted by using scores of preparatory class speaking test and scores of departmental speaking test. It was found that the preparatory speaking test was unable to predict the performances of the test takers in the departmental speaking exam ($r_p(70) = .12, p = .36$). In this respect, it might be concluded that the exam did not have any predictive validity. Moreover, in order to inspect the construct validity of the exam, Pearson product moment correlation coefficients between the students' speaking exam scores and the scores of the students' score in each subtest (*Listening, Structure, Reading and Writing*) of the TOEFL exam were calculated. The results of correlation coefficients were $r_p(70) = .20, p = .10$, $r_p(70) = .04, p = .73$, $r_p(70) = -.08, p = .52$, $r_p(70) = -.62, p = .61$, $r_p(70) = .09, p = .46$ respectively. It is clear from the result that the correlations between the speaking exam scores and the other subtests of the TOEFL-IBT and the total score of the TOEFL-IBT are very low. Based on the results it could be deduced that, there is not a construct validity of the test; probably because, in the study, speaking exam scores were correlated with totally different

constructs, therefore such a result was quite usual. Moreover, correlating results of one examination with another is executed for checking concurrent validity rather than construct validity. It was found out that the inter-rater reliability of the test was not found to be satisfactory because the inter-rater reliability of one pair was found relatively low. As indicated in the table, the correlation coefficients of the first two pairs are $r_p(19) = .91, p < .01$; however, for the third pair $r_p(19) = .49, p = .04$ which is quite low though it is significant. In order to find intra-rater reliability of the raters, correlations the first and the second rating of a rater for the same performance is inspected. The correlation coefficients obtained for the first, second, fourth and the fifth raters are $r_p(19) > .71, p < .01$ which are highly reliable. For the third and sixth raters are $r_p(19) > .56, p = .02$ and $r_p(19) > .58, p = .01$ respectively which are accepted as very low reliable.

Checkpoint which is owned and operated by Latitude Aviation English Services Limited (UK) is an English Language Proficiency (ELP) test developed to test the English language skills necessary for successful English-medium aviation training. It was developed by aviation English testing experts. The examination tests reading, listening and speaking skills of the test takers. The speaking tasks are about aeronautical mechanisms and processes, events in aviation operations and one's future career in aviation and about the aviation industry in general. In the test, fluency, structural and lexical competence, functional competence, including the language of description, sequencing, linking of events, relationships between things, and strategic competence including the ability of the candidate to process visual representations of mechanisms and processes common in aviation with which they may or may not be familiar and to select the most important features of the animation and plan, organize and execute a response based on their linguistic resource are assessed. Rater training and standardization is carried out in three phases. The first phase is pre-standardization. Raters familiarize themselves with test format, tasks, rubric, and

exemplar performances. In the second phase which is standardization, the raters discuss with team leader how to interpret rating scale and procedures for rating, and then listen, rate and discuss a variety of series of exemplar performances. Finally, in the certification phase, the raters rate five samples of speech and the team leader reports a feedback based on the scores given about consistency and severity. Rater reliability is analyzed by using Many Facet Rasch Analysis programme and only the rater who achieve infit mean square values of between 0.50 and 1.50 and, where applicable, outfit mean square values of between -2 and +2. are accepted as raters. In a study conducted by Hamill (2016, as cited in Latitude Aviation English Services, 2017), it was intended to compare the predictive validity of decision scores, which is the lowest score in any part of the test, and composite scores, which is weighting performance in all parts of a test in a single score. Checkpoint scores and scores in Theoretical Knowledge Exam (TKE) which is only given to those who score high on Checkpoint exam are used for analyzing correlations. The results of the study revealed that scores on Checkpoint exam predicted validly the linguistic success in aviation training. Interviewees emphasized that content of Checkpoint was including field related content which they faced in their forthcoming education. The researcher reported that correlations between Checkpoint composite and decision scores and scores on TKE were non-significant and explained the reasons for that a strong relation is unlikely between scores on the two tests because they totally represent different constructs and also the success in aviation training requires being proficient when being instructed in target language; however, involves considerably more than ELP alone because ELP is only one among many variables that affect academic success. In the study, there is not any result of statistical analysis provided related to construct validity of the exam.

In September 2008, a management team at the University of Copenhagen emphasized the necessity for an assessment procedure to endorse university lecturers' skills of English

language use who are supposed to teach at graduate programs at that university. In other words, the test is intended to be developed which assesses if the teachers have a sufficient oral proficiency level for giving lecture and having interaction with graduate students in a university context. Furthermore, when teachers have below the adequate level of use of English language to pass the certification process, the results of the test would provide some feedback about the areas they need to study to be able to be an instructor at those graduate programs. In this respect, the construct of the exam is set if a test taker is able to handle a variety of communicative tasks which are exclusive for university graduate programmes, such as presenting highly complex content material, clarifying, paraphrasing concepts and major points, asking question, understanding and responding to questions from learners, dealing with unclear questions and misunderstandings and negotiating the meaning when necessary. Also, assessment criteria include the generic skills such as fluency, pronunciation, vocabulary, grammar and interaction. The testing procedure consists of three parts; the first one is a warm up; second one is a mini-lecture; and third one is a question and answer session. Only the second and the third part are assessed. As part of the lecture, the task aims to elicit if the examiner is able to give instructions for classroom activities such as pair work or a group work and if the examiner has the necessary language use skills to handle questions of the participants. The third part contains of a question and answer session. The aim of this task is to mock up student/teacher interaction in a classroom context. The criteria are generic criteria and descriptors are at five levels borrowed from the CEFR. During the assessment, analytic criteria is used, each of the participants are assessed immediately by the two examiners separately and is given a global result. The participants are also rated between one and five for generic criteria. After reaching an independent rating, the examiners debate on their ratings and they need to have an agreement for both overall global result and the discrete generic criteria. In cases of disagreement, assessment of an additional examiner is required. The

participants receive a global result and narrative feedback. It is unknown whether there is a validity and reliability study conducted for the speaking test; however, this study could be considered as a successful example for how to develop a speaking test consisting how to define the construct, designing tasks and figuring out the assessment criteria.

In this study, researchers Shatrova, Mullings, Blažejová and Üstünel (2017) discuss the format of a speaking exam developed for the School of Foreign Languages at a Turkish university. The reason for them to develop such a test was the former speaking test's being unsatisfactory in meeting the expectations of stakeholders which causes criticism. Therefore, a new speaking test that measures the abilities of learners to use language appropriately in general social and academic contexts is aimed to be developed. In this respect, the new format of the exam consisted of tasks of conversation with interlocutor and speaking on a topic. In the conversation, interlocutors initiate a conversation with the test taker on everyday issues and the aim is to elicit extended responses rather than plain answers. On the other hand, speaking on a topic includes both picture description covering a wide range of themes and issues and a monologue speech which is an explanation of a topic. For the assessment a holistic rubric was used which consist of five scales ranging from 0-1/ Basic-No response to 4.5 – 5/ Exceeds Expectations which was developed by speaking unit team in the faculty. As a result of subsequent feedback, some amendments were done for bettering instructions and pictures for photo description. All the participants who are assessors and interlocutors as well as students reported positively on the flow and comfort of the exam. In addition, although majority of the assessors displayed consistency in scoring, there were some outliers who were later taken for additional training prior to the exams. Whether there is a reliability and validity analysis carried out about the exam is unknown.

Based on the relevant research above, it could be summarized that when developing a speaking test, the construct should be defined firstly. Then, appropriate tasks should be

selected that would elicit the desired outcome. The process of developing a speaking test continues with defining assessment criteria, rater training to sustain consistency in scoring and procedures to execute the speaking test. In order to judge whether a speaking test is successfully developed could be decided through statistical analysis as explained in the researches above. In order to find construct validity, factor analysis could be conducted to see if sub-skills scored by examiners such as grammar, fluency etc., constitute a construct together. Beside factor analysis, Pearson Correlation or Spearman Correlation analysis could also be used to see if there is correlation among test items and between test items and total score in order to investigate construct validity. In order to find predictive validity, again Pearson Correlation analysis could be used. The correlation between the newly developed test and a future test indicates how much the newly developed test is able to predict the success of a test taker in that future test. Reliability analysis of the test could be investigated through Cronbach Alpha and Pearson Correlation computations. In this purpose, in the researches above, Pearson Correlation Coefficient calculation for inter-rater reliability and computation of Cronbach Alpha for would demonstrate intra rater reliability and the reliability of the test as a whole. These statistical calculations are also approved by Fulcher and Davidson (2007). If there is problem after these calculations, the necessary amendments could be done and the test could be used again.

Chapter III: Methodology

In this study, it is aimed to develop a valid and a reliable speaking test. Also, in this study procedures to execute a speaking test and the consistency in rater behaviors are also aimed to be sustained. In doing so, this test is expected to be a model for the ones interested in developing a speaking test. Therefore, the research questions are in this study as follows:

1. How valid is the test?
 - 1a. Does the test have content validity?
 - 1b. Does the test have construct validity?
2. How reliable is the test?
 - 2a. Is there inter-rater reliability?
 - 2b. Is there intra-rater reliability?
3. What are the views of administrators, teachers and students on the speaking test?

In this study, criterion related validity is not decided to be investigated. This is because, finding another a valid and reliable speaking exam measuring CEFR B1 level and finding trained raters and getting all 164 test takers to have that exam in order to inspect correlation between the scores of the two test seems impractical to research concurrent validity as the research population is very high. Besides, there is no claim about a prospective success of a student based on the result s/he gets from the speaking test; therefore predictive validity of the test is not going to be investigated.

When developing such a test, firstly test specifications are decided to be defined. Model of modular specification suggested by Luoma (2004) is adopted as its being a compact and a more organized method. Quantitative research method is going to be used for conducting validity and reliability calculations of the test. On the other hand, qualitative method is going to be utilized to be able to find views of administrators, teachers and students

about the speaking test. This kind of mixed method approach is also approved by Miles and Huberman (1994) and Strauss and Corbin (1997).

Participants

160 male and 4 female cadet test takers were going to have the exam. Participants had taken 840 hours of English as a foreign language lesson throughout 28-week-long education year. Each of the participants had different backgrounds of English as a foreign language; however, they were accepted at beginner proficiency level since they failed English as a foreign language exemption exam which was held at the beginning of education year.

All of the teachers were supposed to be examiners so as to increase commissions and decrease the number of test takers registered to each commission. Academic background of the examiners is provided at the Table 1 below. There were also three administrators in this research, one of them was head of foreign languages department, the other one was head of preparation class and the last one was head of testing office who were going to take part only in the interview part of the study.

Table 1
Academic Background Profile of Examiner

	Gender		Degree			Experience in Teaching (Year)			Previous Training in Testing Speaking		Experience in Testing Speaking	
	<i>Male</i>	<i>Female</i>	<i>BA</i>	<i>MA</i>	<i>PhD</i>	<i>1-5</i>	<i>6-10</i>	<i>+10</i>	<i>Yes</i>	<i>No</i>	<i>Yes</i>	<i>No</i>
Examiners	16	20	30	3	3	29	2	5	4	32	7	29
TOTAL	36		36			36			36		36	

Data Collection Instruments

Test takers' scores were used for conducting quantitative analysis. Thus, scores were obtained through document used for noting the scores down (See Appendix A and B). In addition to the scores, views of administrators, experienced teachers, inexperienced teachers and students were used for conducting qualitative analysis. Experienced and the inexperienced teachers were selected based on experience in testing speaking. Therefore, an interview was to be carried out to obtain data about the views. Which questions to be asked to which and how many participants are depicted below in table 2. Lastly, to investigate content validity, a comparison table was used for demonstrating data related to the content and abilities in both exam and in the course book.

Table 2
The Questions to be asked to Participants in the Interview

Questions	Administrators	Experienced Teachers	Inexperienced Teachers	Successful Students	Low-successful Students
How effective were the tasks in representing the real world contexts?	X	X	X		
Are the any tasks you would like to add to/extract from the test to make the test more relevant to real world task? Why?Why not?	X	X	X		
Is the number of test takers per commission enough to conduct the test efficiently?	X	X	X		
Are the materials cost-effective?	X				
Is the duration per task long enough to make an effective assessment?		X	X		
Has the speaking examination caused a change in your classroom activities during the lessons? How?		X	X		
Do you think being tested on speaking skill caused you to study more for improving speaking skills? How?				X	X

Data Analysis

This study involves both qualitative and quantitative data. For analyzing quantitative data, SPSS 22.0 is going to be used. In this respect, Cronbach Alpha analysis was used for investigating intra-reliability, and Spearman correlation analysis was used for inspecting inter-rater analysis. Also, an online website www.socscistatistics.com was used for calculating the standard error of measurement to check the reliability of the test. In order to investigate validity of the speaking test, factor analysis was utilized. The qualitative data obtained through interview was to be analyzed by focusing on commonly given answers and lastly data obtained through comparison table was to be analyzed by comparing the abilities to be tested in the speaking test and abilities presented in course book curriculum.

Development Process of Speaking Test

The construct of the speaking test. The aim of this test was to assess the test takers' ability to use target language which is English effectively with regards to CEFR B1 proficiency level. They were expected to make a coherent and through explanations on familiar topics, and also carry out a discussion in an interaction.

- Ability to use language effectively was reflected in;
 - intelligible pronunciation,
 - knowledge of appropriate vocabulary,
 - ability to maintain the flow of speech, despite some failure
 - sufficiently accurate and appropriate knowledge of grammar.
- Ability to make a coherent explanation was reflected in;
 - the ability to connect sentences with linkers, conjunctions and cohesive devices,
 - the ability to make a smooth transition between ideas.

- Ability to make a through explanation was reflected in;
 - relevant knowledge about topic,
 - expression of opinion about topic,
 - expression of reasons for justification,
 - wrapping up essential points to come to a conclusion.
- Familiar topics were which test takers might have some knowledge or experience such as (Cambridge PET, 2016, p. 68);
 - sport,
 - environment,
 - daily life,
 - personal opinions, experiences and feelings,
 - hobbies and leisure,
 - services,
 - social interaction,
 - travel and holidays etc.
- Ability to carry out a discussion in an interaction was reflected in;
 - participating in discussion,
 - taking turns appropriately,
 - asking and expressing reasons,
 - leading a discussion to develop interaction,
 - negotiating the ideas to compromise.

Selection of the test tasks. To be able test the test takers' speaking abilities described in the construct, two tasks, a sustained monologue production task and a pair interaction task, were decided to be selected. To conduct sustained monologue production task, topic

explanation activity was prepared. In this respect, a topic pool is created. The test taker was supposed to talk about one of the topics from the pool. Then, the test taker was asked one additional question to elicit further talk on that topic. Moreover, in order to conduct pair interaction task, a discussion activity was prepared. In this discussion, test takers as a pair discuss several options to solve a problem together and come to a conclusion. A sample for each speaking task is presented in appendix C and D.

Assessment criteria. Assessment rubric was designed in the light of construct and speaking tasks of the speaking examinations. In this respect, since each task contained unique features, two separate assessment rubrics were decided to be used. In this examination, an analytic, multi-trait and generic rubric was decided to be used because of the aforementioned advantages of them. However, there seemed to be only one drawback of using an analytic rubric which was low inter-rater reliability. In order to eliminate or at least minimize the effect of this drawback, rater training was planned which is to be explained in another section below. Descriptors of the rubric were defined by firstly analyzing the existing rubrics. In this respect, the most efficient assessment rubric was decided to be of Cambridge English Preliminary (PET) examination which makes assessment of language proficiency in four skills one of which is speaking skill in CEFR B1 level. Nevertheless, that rubric lacked the feature of assessing sustained monologue production in terms of fluency. Therefore, being inspired from CEFR documents, fluency was added as a criterion. In order to assess speech performance for the first task which was sustained monologue production, the rubric consisted of grammar, vocabulary, pronunciation, fluency and production as criteria. On the other hand, to assess speech performance for the second task which was paired interaction, the rubric consisted of grammar, vocabulary, pronunciation and lastly interaction. The reason for not having fluency is going to be explained in piloting phase of the examination.

Cut scores were already defined in the original version of the assessment rubric; however, because of the testing and evaluation criteria of the institution, those cut scores had to be adapted accordingly. Therefore, each of the tasks was decided to worth 50 points and because of that each of the five sub-components of the rubric for monologue production task is updated as 10 points at maximum. For each sub-component, good level of performance was worth 10-9 points, between good and average performance was worth 8-7 points, average performance was worth 6-5 points, between average and poor performance was worth 4-3 and poor performance was worth 2-0 points.

On the other hand, since pair interaction task consists of four sub-components, each of them does not have integer value. Thus, one of the sub-components is decided to outweigh the others which is also supported by Luoma (2007). In this respect, regarding that it was an interaction task, the sub-component of “interaction” of the rubric was decided to outweigh the others. The interaction was worth 20 points at maximum, and the other each sub-component was worth 10 points. Except for interaction, for each sub-component, good level of performance was worth 10-9 points, between good and average performance was worth 8-7 points, average performance is worth 6-5 points, between average and poor performance was worth 4-3 and poor performance was worth 2-0 points. For the interaction sub-component, good level of performance was worth 20-17 points, between good and average performance was worth 16-15 points, average performance was worth 14-11 points, between average and poor performance was worth 10-8 and poor performance was worth 7-0 points. Each speaking assessment rubric was presented in appendix E and F. To count a performance as CEFR B1 level; minimum score should be 60 out of 100.

Procedures to conduct speaking test. This speaking test aimed to test speaking abilities of the test takers at CEFR B1 level. The test was comprised of two different tasks each of which had 50 points of weighting. The first task lasted for three minutes totally; one

minute for preparation and two minutes for speech delivery. The second task lasted for four minutes totally; one minute for preparation and three minutes for pair discussion.

Pairing test takers for the second task was also planned in the basis of studies conducted to find out which type of pairing would be more productive. According to Norton (2005), if one of the test takers is more proficient than the other, or if both partners know each other, they could perform higher linguistic abilities. Therefore, before pairing test takers, mean scores of the previous tests were analyzed and then test takers having the highest scores were matched with the ones relatively lower proficient ones.

In order to carry out the test successfully, there were some documents that examiners must have. Speaking commissions were made up of two examiners who are teachers of the students. They were given speaking examination file. The file consisted of a list of test takers registered to have the test on that commission, document for noting down speaking scores individually for each of the examiner, document for noting down total scores of each examiner and assigning the average score, assessment rubrics for each task, five different samples for each task, and numbers from one to five for drawing lots to decide which sample of the task the test taker was going to have. Test takers were registered to commission which had teachers as examiners who were not teaching them at that time.

It is important for examiners to know how to conduct the examination. Speaking examination started with a warm up phase in which an examiner asked several questions about the test taker's feelings, the weather etc. to familiarize him or her for the test. This part was not assessed. Then, test taker picked a number from the poll to decide the sample task for the first task. The test taker was given the instruction and started with the first task. After test taker finished the first task, his/her pair was called to take the first task. When that partner finished the first task in the same way, both test takers moved on to the second task. Similar to

the first task, one of the test takers picked a number from the poll to decide the sample task for the second task and instruction was given for both partners. Then, through interaction, they undertake the task together as pairs.

Assessment was done during the performances of the test takers. Test takers were assessed on their individual performances for both the first task and the second task separately. At the end of the test, each rater assigned a score for each sub-skill and added them up to have the final score for the first task and did the same for the second task. Then, each rater added the score for the first task and the second task to have a final score for the whole speaking performance. Lastly, each final score of two raters was averaged to define one single final score of the whole performance of the test taker.

Rater training. In order to sustain a consistency among raters in terms of scoring performances and carrying out the speaking test successfully, a workshop was planned to train raters.

Firstly, theoretical information was given about speaking examination. In this respect, a presentation was prepared in which there was information about the profile of test takers, learning outcomes of speaking skill based on CEFR B1 level, the construct of the speaking exam, speaking exam tasks, assessment criteria, documents for noting down scores, procedures to undertake the exam and videos of good, average and poor sample performances which were taken from Cambridge English's YouTube channel, named *Cambridge English: Preliminary for Schools, Victoria and Chiara*, and *Cambridge English:KEY Speaking*.

Chiara's performance was regarded as good performance which deserved a score in band A in terms of all of the sub-skills. Thus, a good performance in grammar was to utter mostly accurate speech and employ some complex forms as well, in vocabulary was to use a variety of relevant lexis accurately, in pronunciation, was to sound almost always clear,

paying attention to stress and intonation, in fluency, was despite the evident gaps, flow of the speech did not tire listener, in production, was to produce an extended discourse with satisfactorily relevant information about the topic by including cohesive devices appropriately and accurately, and lastly in interaction, was to respond and initiate the interaction without any support.

On the other hand, Victoria's performance was regarded as an average performance which deserved a score in band B in terms of all of the sub-skills. Thus an average performance in grammar was to show good command on simple grammatical forms, in vocabulary was to use words repetitively, in pronunciation, was to sound intelligible despite L1 problems in stress and intonation, in fluency was false starts and apparent reformulations, in production was to produce an extended discourse with adequate information about the topic by including basic cohesive devices, in interaction was in generally responding position during the interaction.

Finally, in order to show a poor performance of speaking in terms of grammar and vocabulary sub-skills, a KET examination video was watched. In that video, Luis' performance was decided to be a poor one in terms of grammar and vocabulary. Thus, a poor performance in grammar was to display only limited control of some simple structures using isolated phrases, and in vocabulary, was to use limited range of vocabulary and being based on a few different words. Poor performances for pronunciation, fluency, production and interaction were defined independently from the sample videos. It was decided that a poor performance in pronunciation was a heavy influence of L1 accent which causes unintelligibility, in fluency was a huge silence between utterances that distracts listeners, in production was to produce speech at sentence level with scarce information about the topic, not to have cohesive devices, in interaction was to be unable to respond and contribute without help during the interaction.

After theoretical information was given to raters, they were asked to implement a mock examination to carry out all procedures and rating. Therefore, the test was piloted and feedback was asked to make the required amendments.

Piloting of the speaking test. The test was piloted by all of the raters who were the instructors at school. The problems aroused by raters could be categorized into two which were problems related to procedures of the exam and scoring.

One of the problems related to procedures of the exam was the duration of the tasks. The first task was planned to last for one minute for preparation and one minute for delivering the speech. Raters objected this issue claiming that it would be insufficient time for delivering a speech resulting in a pressure on test takers. Therefore, one minute more is added for speech time. In addition to duration of the task, it was asked by the raters whether it was available to help by asking a question if test taker is stuck. In order to sustain a standardization, it was decided that no matter a test taker was stuck or not, he or she was going to be asked a question either when they were stuck or finish the speech. Also, if test takers were unable to understand anything in question, an explanation could be made.

One of the problems related to scoring test taker's performance was related to assessment rubric for the second task. In this task, test takers are supposed to discuss several options to solve a problem and then come to a solution. It was observed by raters that during the execution of second task, test takers generally produced a sentence level talk or a very little extended talk because the nature of the interaction required a frequent turn taking. Therefore, they claimed that it was too difficult to assess fluency. As a result, fluency was excluded in this assessment rubric. Another problem issued by the raters was that, they were unsure to assign which score to test takers' performance because the assessment criteria were too demanding in terms of scoring as being an analytic rubric. To overcome this problem,

raters were recommended assess performance firstly according to the band it was appropriate, and then assign a score in that band. Moreover, they were advised assign a band performance in the first 15-20 seconds for sub-skills which were distinctively good or poor, except for “production”. If not possible, they were advised focus on some sub-skills more to assign a band performance in the second 15-20 seconds. Once they had made a decision on band performance of some sub-skills, they were suggested focus on the other sub-skills. Lastly, once they had decided on band performances for all sub-skills, they are advised observe fluctuations in the performance and make the necessary changes.

After analyzing the correlations of the scores, the commissions, raters in the commission in which there was not a significant positive correlation between the raters were called back to revise their training with a more experienced rater. In that session, they discussed their scoring, and then in the leadership of experienced rater, they were taught how they should have approached to give a right score.

Chapter IV: Findings

RQ1: How Valid Is The Test?

1a. Does the test have content validity? Validity issue is a broad term and thus to count a test a valid test, it is necessary to inspect a test from different dimensions. Therefore, as stated above, content validity and construct validity of the test is going to be identified. In order to find an answer to RQ1a, the table 3 below is displayed in which aims and outcomes in the curriculum of the course book are compared to abilities tested thorough the test tasks.

Table 3

A Comparison Table of Abilities both Tested in the Exam and Taught in the Course

	Abilities Tested in The Exam	Course Book* Speaking Curriculum
Linguistic abilities tested in both tasks	Intelligible pronunciation, intonation and stress	Practicing varying intonation to convey your attitude Asking questions and making statements with correct intonation Answering tag questions using proper grammar and intonation to accurately express what you think
	Sufficiently accurate and appropriate knowledge of grammar	Using modals to express obligation, prohibition and recommendation Adding tag questions to find out what someone thinks Answering tag questions using proper grammar and intonation to accurately express what you think
	Knowledge of appropriate vocabulary	Using adjectives, fixed phrases and idioms to express emotions
	The ability to connect sentences with linkers, conjunctions and cohesive devices, Making a smooth transition between ideas	Using sequence expressions to clarify order of events
Abilities tested in sustained monologue production task	Relevant knowledge about topic	Giving advice, describing a situation
	Expression of ideas about topic	Clearly introducing the topics
	Providing reasons for justification	Giving reasons and examples to support opinions Explaining reasons to justify statements about personal preferences
	Expressing essential points to conclude	Using summary or recap techniques

Abilities tested in paired interaction task	Participating in discussion	Participating in a group discussion Asking for clarification, Clarifying what you say Adding tag questions to find out what someone thinks
	Taking turns appropriately	Taking turns to make a conversation go smoothly
	Asking and expressing reasons	Asking for reasons Giving reasons and examples to support opinions
	Managing a discussion to develop interaction	Leading a discussion
	Negotiating the ideas to compromise	Using expressions to introduce statements of agreement and disagreement

*Oxford Q Skills for Success Listening and Speaking 3

Table 3 above depicts the matching abilities between the ones tested in the speaking exam and taught through the speaking courses during the education year. There are 15 abilities decided to be tested in the exam totally which are also presented in the section of “methodology – the construct of the speaking test”. In the course book curriculum “Q: Skills for Success, Listening and Speaking 3” there are 27 different speaking abilities included to be taught. 14 of the abilities to be tested match with 19 of the abilities included in the curriculum. It could be understood that 70% of the abilities in the course book curriculum are tested in the exam directly. There is an ability in the course book curriculum tested indirectly which is “Using verb contractions to increase naturalness of speech.” It is not tested directly because it is not evaluated as an ability which a CEFR B1 level student should necessarily demonstrate; however, it affects intelligibility and fluency, thus a test taker’s performance is affected as well.

There are some speaking abilities in the course book curriculum which are tested neither directly nor indirectly in the speaking examination such as, implying opinions to avoid stating them too directly, to conduct/participate a survey, include time for questions after a presentation, using T chart to take notes for a presentation, make notes to give

presentation/interview, discuss with a partner, attitudes about the relationship between money and happiness and prepare a dialogue with a partner to improve your conversational skills. There is an ability tested in the exam although it is not included in the curriculum which is the ability “to maintain the flow of speech, despite some failure”.

Apart from the abilities tested in the speaking exam, topics about which test takers have to speak in sustained monologue production task are thought to hold similarities with the course book content. It could be regarded that the more there is similarity between the test task topics and course book content, the more the familiarity of test takers would be. Thus, content validity of the speaking exam could be considered to be sustained more strongly by doing so. Table 4 depicts the topics about which test takers are supposed to talk in the exam and which course book content includes.

Table 4
Comparison of Topics Tested in the Exam and Discussed in Classroom Sessions

Speaking Test (Sustained Monologue Production)		Course Book Content	
Personal Preferences	Do you prefer reading book from tablet/kindle or reading the hardcopy of the book? Explain your ideas by giving reasons and examples.	Is change good or bad?	New Perspectives
Health	What is the importance of health? What can you do to stay healthy? Explain your ideas by giving reasons and examples.	What is more important; taste or nutrition?	Food and Taste
Purpose in Life	What is the importance of having a goal in life? What is your biggest goal in life? How do you plan to achieve that goal? Talk about the goals that you’ve achieved and how you did it.	What can we learn from success and failure?	Success
Technology	What are the pros and cons of using the internet? Explain your ideas by giving reasons and examples.	Do we need technology to communicate	Keeping in touch

	What do you use a computer for? How would life be without computers? What are the good and bad things about having a computer?	long distance?	
Social Life	Friendship is the most important relationship; do you agree or disagree? Why? How do you maintain a good friendship? What things should friends never do? What is it like to meet new people? What kind of people do you like to meet? What impressions would you like leave on the people that you meet? Explain your ideas by giving reasons and examples.	Are first impressions accurate?	First Impression

Table 4 above depicts the matching topics between the ones to be talked about in the speaking exam and discussed through the speaking courses during the education year. There are 14 topic questions decided to be tested in the exam totally. In the course book curriculum “Q: Skills for Success, Listening and Speaking 3” there are 10 different speaking topics included in the course book content. 5 out of 14 topic questions match with 5 out of 10 topics included in the course book content. It could be understood that 50% of the topics in the course book content are tested in the exam directly.

1b. Does the test have construct validity? Construct validity of a test carries great importance, as mentioned earlier in this study; it forms the basis of any type of validity (Mislevy, 2007). From the ways that are presented by Cronbach and Meehl (1955), factor analysis (FA) is preferred as its being widely used analysis in test scores (Fulcher, 2003). During the examination, test takers are evaluated on their performances in two tasks in terms of vocabulary, pronunciation, grammar for each task, fluency and production only for one task and interaction only for the other task. Thus, these nine sub-skills constitute a total speaking score of a test taker. Before conducting FA, the factorability of the nine sub-skills that constitute the total speaking score is examined. In this respect, firstly, the correlation

values among sub-skills in the exam are examined and they indicate that all the items correlate ($r \geq .54$, $p. \leq .01$) with one another. This result signals the possibility of reasonable factorability. At the next step, the results show that the KMO coefficient was .937 which could be considered as a great value to administer FA because a value of .60 or greater is regarded to be sufficient (Razı, 2012). Lastly, since the result of communalities was above .3 for each sub—skill, all those sub-skills are assumed to contain common variance with the other ones. In conclusion, it could be decided that pre-analysis results indicate FA could be a suitable technique to be conducted for investigating construct validity.

In order to decide how many components there should be, several criteria are taken into account. In this respect, only the components which have eigen-value of 1.0 or above are allowed to remain (Pallant, 2001). Moreover, Büyüköztürk (2007) highlights that the items under the same factor are expected to have high factor loads ($\geq .450$), and lesser ones should be eliminated. In addition, he continues that items are not permitted to have factor loads under two different factors; therefore, a difference needs to be between the two highest factor loads of at least .100. Furthermore, as suggested by Pallant (2001), items which do not load on any factor should also be eliminated. Bearing the criteria in mind, principal component analysis revealed one component with eigen-value of 6.42 explaining 72.1% of the variance. Result of scree plot is displayed in Figure 6 below. In table 5 below results of factor analysis are presented.

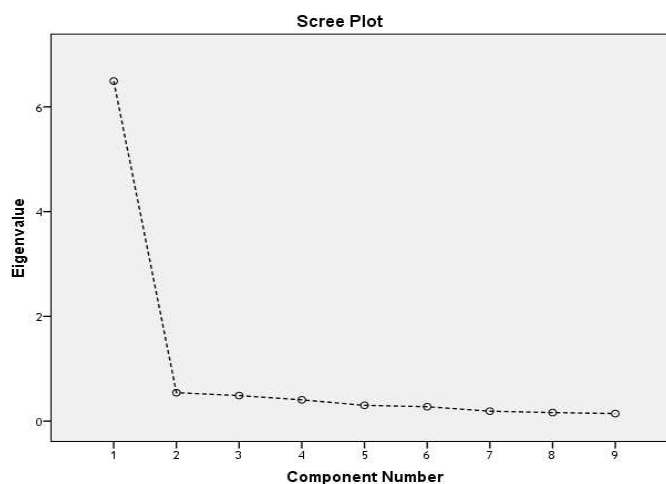


Figure 6. Scree plot.

Table 5

Mean Values, Communalities and Factor Loadings Based on PCA for 9 Sub-skills of Speaking Test (N=164)

Items/Components	Mean Values(SD)	Communalities	Speaking Score
Task 1 Grammar	6.80 (1.32)	.662	.918
Task 1 Vocabulary	7.27 (1.48)	.843	.896
Task 1 Pronunciation	7.40 (1.31)	.737	.867
Task 1 Fluency	7.35 (1.63)	.731	.858
Task 1 Production	7.54 (1.50)	.803	.855
Task 2 Grammar	6.83 (1.38)	.696	.834
Task 2 Vocabulary	7.35 (1.51)	.752	.828
Task 2 Pronunciation	7.38 (1.25)	.686	.813
Task 2 Interaction	15.41 (2.90)	.583	.763

Explained variance: Total = 72.1%; Factor 1 = 72.1%.

The only component is named as Speaking Score because these nine sub-skills constitute the overall speaking ability of a test taker. Following FA, internal-consistency reliability is computed and Cronbach's alpha coefficients are revealed for the nine sub-skills in the total speaking score ($\alpha = .93$), therefore, the Cronbach's alpha coefficients provide that the overall scale has very high internal consistency.

RQ2: How Reliable Is The Test?

2a. Does the test have inter-rater reliability? In the development phase of the study a rater training session is conducted. During the session, raters are instructed on how to apply the scale appropriately. That is considered to be important in providing consistency between the raters. In doing so, similar scores from different raters for the same performance are expected to occur.

There are 18 commissions in the exam and for each commission there are maximum 10, minimum 8 test takers. Since the number of test taker per commission is very low, non-parametric test is used (Pallant, 2001). Spearman rank order correlation analysis is conducted to investigate inter-rater reliability. The same technique is used for each commission. In the table 6 below, inter-rater reliability of each commission is presented.

Table 6
Spearman Rank Order Correlation Analysis between the 1st Rater and the 2nd Rater

Commissions	Raters	r_s
1 st *	1 st & 2 nd Rater	.93
2 nd *	1 st & 2 nd Rater	.88
3 rd *	1 st & 2 nd Rater	.88
4 th *	1 st & 2 nd Rater	.29
5 th *	1 st & 2 nd Rater	.99
6 th *	1 st & 2 nd Rater	.85
7 th *	1 st & 2 nd Rater	.97
8 th **	1 st & 2 nd Rater	.92
9 th **	1 st & 2 nd Rater	.79
10 th *	1 st & 2 nd Rater	.89
11 th **	1 st & 2 nd Rater	.79
12 th **	1 st & 2 nd Rater	.88
13 th **	1 st & 2 nd Rater	.69
14 th **	1 st & 2 nd Rater	.74
15 th **	1 st & 2 nd Rater	.81
16 th **	1 st & 2 nd Rater	.76
17 th **	1 st & 2 nd Rater	.88
18 th **	1 st & 2 nd Rater	.97

** There are 10 test takers in the commission.

* There are 8 test takers in the commission.

In the examination, in 17 out of 18 commissions, raters' ratings correlate positively and strongly with each other ($r_s \geq .69$, $p \leq .03$). It could be deduced that assessment perception of raters might display resemblances upon considering the consistency occurred after speaking examination. Also, bearing the Butler's (1985) and Cronbach's (1990) statements above in mind, the correlation values in these commissions could be accepted as good values in terms of inter-rater reliability.

On the other hand, although 1st and the 2nd raters' total scores are positively correlated in the 4th commission, the result is insignificant ($r_s(8) = .29$, $p > .05$). It could be deduced that, the correlation in this commission is random and unreliable.

2b. Does the test have intra-rater reliability? After investigating the inter-rater reliability, it is also important to check intra-rater reliability. As H.D. Brown (2004) indicates, intra-rater reliability displays how consistent a rater with himself/herself in terms of scoring. In the table 7 below, results of intra-rater reliability analysis of each rater in each commission is displayed.

Table 7

Intra-rater reliability of the 1st and the 2nd Rater (n of Items =9)

Commissions	Raters	Cronbach's Alpha
1 st	1 st & 2 nd Rater	.97 & .94
2 nd	1 st & 2 nd Rater	.91 & .91
3 rd	1 st & 2 nd Rater	.96 & .90
4 th	1 st & 2 nd Rater	.30 & .11
5 th	1 st & 2 nd Rater	.72 & .88
6 th	1 st & 2 nd Rater	.90 & .94
7 th	1 st & 2 nd Rater	.69 & .87
8 th	1 st & 2 nd Rater	.84 & .78
9 th	1 st & 2 nd Rater	.76 & .81
10 th	1 st & 2 nd Rater	.96 & .95
11 th	1 st & 2 nd Rater	.90 & .69
12 th	1 st & 2 nd Rater	.84 & .84

13 th	1 st & 2 nd Rater	.80 & .87
14 th	1 st & 2 nd Rater	.87 & .61
15 th	1 st & 2 nd Rater	.77 & .93
16 th	1 st & 2 nd Rater	.92 & .88
17 th	1 st & 2 nd Rater	.95 & .91
18 th	1 st & 2 nd Rater	.95 & .92

The scores of the test takers on the 9 items (Task 1 Grammar, Task 1 Vocabulary, Task 1 Pronunciation, Task 1 Fluency, Task 1 Production, Task 2 Grammar, Task 2 Vocabulary, Task 2 Pronunciation and Task 2 Interactive Communication) are then analyzed to find out the intra-rater reliability of the raters in the speaking test. Reliability analysis revealed a Cronbach's alpha score of $\alpha \geq .70$ over 9 items in the speaking test for 33 raters out of 36. These scores indicate that vast majority of raters have acceptable intra-rater reliability. On the other hand, the second rater in the 14th commission with Cronbach's alpha score of $\alpha = .61$ over 9 items has a weak intra-rater reliability and raters in 4th commission with Cronbach's alpha score of $\alpha = .30$ over 9 items and $\alpha = .12$ over 9 items respectively do not have intra-rater reliability at all.

Overall reliability of the speaking test is supposed to depict us how much the speaking test is dependable as a whole. Therefore, 18 items (9 items from the 1st rater plus 9 items from 2nd rater) are analyzed to find out the reliability of the speaking test. Reliability analysis revealed a Cronbach's alpha score of $\alpha = .95$ over 18 items in the speaking test. This score indicates that speaking test is highly reliable.

In addition to Cronbach alpha analysis of reliability, standard error of measurement is also an indicator of reliability (Genesee & Upshur, 1996). As long as standard error of the test is smaller, the reliability of the test is higher (Luoma, 2004). According to the computation on an online platform, standard error of the speaking test is $SE = .095$ (Social Science Statistics,

2019). This score indicates that real score of a test taker might change ± 1 from his/her score assigned by the raters.

RQ3: What Are The Views Of Administrators, Teachers and Students About Speaking Test?

Participants' view on authenticity of the speaking test is investigated through questions of "How effective were the tasks in representing the real world contexts?", and "Are there any tasks you would like to add to/extract from the test to make the test more relevant to real world task? Why? Why not?". By looking at the answers given to the first question which is asked to three administrators, five experienced teachers and five inexperienced teachers, almost all of the participants agreed that the tasks include the skills and abilities required in the real world context and therefore they represent real world context. Answers given by administrators, some of the experienced and some of the inexperienced teachers are provided below.

"These two tasks allow students to communicate in different contexts along with displaying many speaking abilities. For example, in the first task, students are expressing their views on the topics and producing a speech at a discourse level. In the second task, additionally, students are collaborating on a task to come to an agreement by exchanging their opinions. Both of the tasks have their places in real life." Head of Foreign Languages Department.

"Students are tested in both monologue and dialogue. No matter what the topics are, students will enter into this kind of speaking contexts in the real world. Therefore, the exam is very effective in preparing students for the real world task." Inexperienced Teacher 1.

“I think they are testing the abilities which are necessary in the real world context. They are entering into an interaction, exchanging their views, showing whether they agree or disagree by presenting reasons. It is obvious that it is impossible to bring a real life context with its all features in an exam context; however, this exam is successful in terms of projecting a student’s potential in real life contexts.”
Experienced Teacher 1.

However, only one experienced teacher stated that neither this speaking exam nor any other speaking exam can represent real world context. The answer is provided below.

“Well, I don’t believe any exam could represent a real life situation because, although in both contexts the same abilities are necessary, real life situation is always completely different from what the exam is. The exam is quite successful in terms of the testing the abilities, but, I don’t think it prepares a student for the real life.”
Experienced teacher 3.

By looking at the answers given to the second question which is asked to three administrators, five experienced teachers and five inexperienced teachers, all of the administrators, two experienced teacher and two inexperienced teachers agreed that there is no need to extract or add any new task because the current tasks are able to test the abilities to be found in CEFR B1 level and taught in course book. Some of the answers are provided below.

“For the moment, there is no need to add or extract anything because both tasks are able to test the abilities which are also taught throughout the courses.” Head of Preparation Class.

“I think these tasks are OK. There is no need for an extra task.” Inexperienced Teacher 4.

“I think these two tasks are able to test the abilities that we as teachers want from CEFR B1 level students to display. Therefore, there is no need for such a change.” Experienced Teacher 5.

On the other hand, three inexperienced and three experienced teachers stated that either there should be one more new task or sustained monologue task should be changed with a new one. The answers are provided below.

“I think students must be tested in real life environments; they should be assessed in simulations.” Inexperienced teacher 1.

“I think, monologue task in the exam is too old-fashioned and might be changed with a task in which a student talks based on an information, a graph etc. Also, a conversation task also might be added.” Inexperienced teacher 3.

“I think a role play task might be added to see how well they are able to perform duties for certain roles.” Experienced teacher 2.

“I think, an interview could also be added as a task to see how a test taker is able to enter into a conversation.” Experienced teacher 3.

Participants' view on practicality of the speaking test is investigated through questions of “Is the number of test takers per commission enough to conduct the test efficiently?”, “Are the materials cost-effective?” and “Is the duration per task long enough to make an effective assessment?”. By looking at the answers given to the first question which is asked to three administrators, five experienced teachers and five inexperienced teachers, all of the participants agreed that the number of test taker per commission is quite enough to conduct the test efficiently. The answers given by some of the participants are provided below.

“The number was quite enough. No one complained after the test because of the number of test takers per commission and many of the commissions finished examination after one hour plus fifteen minutes.” Head of Testing Office.

“Yes, definitely. We were not tired too much; especially testing students as pairs was so efficient.” Inexperienced Teacher 4.

“Yes, there wasn’t any problem. The exam finished in one hour and fifteen minutes.” Experienced Teacher 5.

By looking at the answers given to the second question which is asked to three administrators, all of the participants agreed that the exam is cost-effective. The answers given by administrators are provided below.

“In this institution, government provides all the expenses and we had enough budget. Since we will reserve these documents for the future, we can say that it is cost effective for us.” Head of Foreign Languages Department.

“We were able to print and copy the documents for free since the institution has the necessary facility and government pay for it. Therefore, it is cost effective for us.” Head of Preparation Class.

“It was cost-effective because government pays for the material. We can also preserve these materials for the future. Therefore, especially in the long term, it is cost-effective.” Head of Testing Office.

By looking at the answers given to the third question which is asked to five inexperienced teachers and five experienced teachers, it could be understood that while inexperienced teachers find it difficult to score all sub-skills while also observing the test taker in such a short time, experienced teachers indicated that they did not experience such a

difficulty in scoring all sub-skills in given time. The answers given by some of the participants are provided below.

“At first it was difficult for me to focus on all sub skills to assess the performance and observing the test taker at the same time. Actually, I had to take my time to look my notes and score the performance after the exam is over.” Inexperienced Teacher 3.

“It was challenging to focus on all the sub skills at the same time. It was hard to discriminate differences between performance descriptors while observing the performance of test taker. In such a time, it was difficult for me to make a fair judgment.” Inexperienced Teacher 5.

“That was OK. Especially, after one minute, weaknesses and strengths of the test takers emerge clearly, therefore, there is no problem in scoring the performance in the given time.” Experienced Teacher 2.

“Absolutely, the given time is both OK for scoring and for test taker to give answer. I had no problem in doing that.” Experienced Teacher 5.

Participants’ view on washback of the speaking test is investigated through questions of “Has the speaking examination caused a change in your classroom activities during the lessons? How?” and “Do you think being tested on speaking skill caused you to study more for improving speaking skills?”. By looking at the answers given to the first question which is asked to five experienced teachers and five inexperienced teachers, almost all of the participants agreed that the speaking examination somehow affected their classroom activities in a positive way by sparing some extra time for speaking activities. The answers given by some of the participants are provided below.

“I’ve always believed speaking skills has great importance. Bearing that there is a speaking exam, I’ve always paid attention to spare some time for speaking activities.”
Inexperienced Teacher 1.

“During the education year, I designed speaking activities which are supposed to appear in speaking exam. I think that worked, because my students thanked me a lot for having them study for those tasks.” Inexperienced Teacher 3.

“I always bring some activities for speaking skills, but speaking examination boosted the motivation both for me and my students.” Inexperienced Teacher 5.

“Not that much, the only change in my classroom activities was involving similar tasks with the ones in the exam.” Experienced Teacher 3.

“For sure, I tried to bring more activities which are similar to the ones in speaking examination. As a conclusion, my students were more successful in speaking examination.” Experienced Teacher 4.

On the other hand, there is only one experienced teacher who declared that this speaking examination did not cause any change at all. The answer given by experienced teacher 2 is provided below.

“Actually, I’ve always been preparing speaking exercises, therefore, it didn’t cause me to change or add anything new in my teaching practices.” Experienced Teacher 2.

By looking at the answers given to the second question which is asked to five successful students and five low successful students, it could be understood that all of the students agreed upon that they were motivated with the existence of such a speaking examination, therefore, they participated a lot more in the classroom activities to improve their speaking skill.

“I always wanted to improve my speaking skills, therefore with the help of this exam, I always tested my skills and tried to improve myself in classroom activities.”
Successful Student 2.

“Yes, during the lessons I did my best to get a good score from the speaking exam.”
Low Successful Student 2.



Chapter V: Discussion

Discussion For RQ 1a And 1b

In this research question it is aimed to investigate validity of the exam in terms of content and construct validity. To achieve this aim, abilities to be tested via speaking tasks in the exam are illustrated in a table and then they are compared to see how much they match with the aims and outcomes in the course book which are also presented in the same table. Similar to the comparison of abilities to be tested, the comparison of the speaking topics in the exam with the content of the course book is carried out in the same way. The idea of such a comparison to determine the content validity of a language test is also approved by Anastasi (1988), Bachman (1999), J. D. Brown (1996), Heaton (1990) and Hughes (1989). As stated above, Genesee and Upshur (1996) points out that views of experts should be taken to judge relevance of the test content and adds there is no statistical way of doing this.

Upon looking at the comparison table above, it could be seen that the abilities tested in the exam match with the vast majority of aims and outcomes of the course book. Besides, speaking topics in the exam, show medium similarity with the content of the course book. In a study conducted by Sak Kıymazaslan (2008), an interview is conducted and participants stated that the purpose of the exam is to measure the speaking abilities of students regardless of their training background; therefore, content of the speaking exam does not necessarily be representative of the content or objectives of language programme. In another study by Zhihong et al. (2016), test specifications and the abilities tested via speaking exam task are compared to investigate content validity. Pragmatic competence, sociolinguistic competence and organizational competence are decided to be represented through tasks of group discussion, pair work and personal statement.

This thesis study shows similarity with the study conducted by Sak K1ymazaslan (2008) in terms of having moderately high degree of content validity. Both of the studies reveal that the speaking exams cover the topics taught in the curriculum up to some extent; since, the both do not necessarily include all of the topics/content of the course book, regarding that the utmost aim is to test proficiency. On the other hand, the study conducted by Sak K1ymazaslan (2008) disapproves of the inclusion of objectives of course book, but this thesis study aimed to include as many objectives/aims and outcomes of the course book as possible because the speaking exam targets to test CEFR B1 level of proficiency and the level of the course book is also CEFR B1. Apart from the difference of covering abilities, the both studies differentiate in collecting data to investigate content validity. While in the study conducted by Sak K1ymazaslan (2008), the content validity is investigated via interview, in this thesis study content validity is investigated through comparing the abilities and topics covered in the exam with the aims and outcomes and content of the course book. The reason why this study has a medium content validity in terms of topic in the speaking exam is that CEFR B1 proficiency level is not limited to only the topics in the course book curriculum and when being too dependent on the course book the authenticity of the speaking exam would fall. Therefore, different topics which are considered to be suitable for CEFR B1 proficiency level are included in the speaking test. On the other hand, the reason for having a medium content validity in the study by Sak K1ymazaslan (2008) could be that the speaking exam is mostly independent from the course book curriculum.

The present study shows similarity with the study conducted by Zhihong et al. (2016) in terms of content validity. Both studies collected data to investigate content validity by using comparison instead of carrying out an interview. Also, in both studies, content validity is reported as high degree in terms of the abilities tested in the exam. However, the both studies have some differences. While in the study by Zhihong et al. (2016) the abilities tested

by tasks are compared with test specifications, in this thesis study, abilities and topics covered by the exam are compared with the ones in the course book curriculum. The comparison of the test specs with the abilities tested by the tasks in the exam is considered to effect the result of such a high content validity in the study of Zhihong et al. (2016) because the tasks in the exam are selected according to the test specifications. Thus, if those abilities could have been compared to content of the course book, a different result could have emerged. As stated above, the reason for having high content validity in this thesis study in term of the abilities tested in the speaking exam could be because the proficiency level of the course book and proficiency level of the speaking exam are same, CEFR B1 level.

In order to investigate the construct validity of the test, statistical analysis of FA is conducted. In this respect, mean scores of each nine sub skills scored by two different raters are taken and decided as the variables that constitute the total speaking score of a test taker. Computing factor analysis to explore construct validity is also approved by Cronbach and Meehl (1955), Anderson et al. (1995), Xiaoqing (2003) and Fulcher (2003).

Upon looking at the results of the FA, as stated above, the construct validity of the exam is sustained. There are several studies in which construct validity of an exam is inspected through statistical analysis. For example, in a study conducted by Zhou (2016) both FA and multi facet Rasch analysis are conducted to investigate construct validity. In addition, while Sak Kıymazaslan (2008) conducted Pearson product moment correlation between speaking exam scores of students and subtest scores of TOEFL IBT test, in a study conducted by Zhihong et al.(2016) scores of the subtests of speaking exam are compared by using Pearson product moment correlation in the same way both with each other and total scores of the speaking exam.

This thesis study has a similarity with the study of Zhou (2016) in terms of investigating construct validity by using FA. According to the results of the both studies, both tests have construct validity. However, studies conducted by Zhihong et al. (2016) and Sak K1ymazaslan (2008), utilized Pearson product moment correlation analysis and therefore differentiate from this thesis study in terms of investigating construct validity. In addition, in the study conducted by Zhou (2016), multifacet rasch analysis is also used for checking construct validity. In this thesis study as well as the studies conducted by Zhihong et.al.(2016) and Zhou (2016), construct validity is established. In these studies, items in the same test are inspected to find inter-item correlations to investigate construct validity through FA and Pearson correlation calculations. Nevertheless, in the study by Sak K1ymazaslan (2008), scores of speaking exam is compared with the subtest scores of a totally different exam, which was TOEFL IBT, through Pearson correlation calculation. As a result, since each of the tests is a very different construct in itself, low and insignificant inter-item correlation occurred in the study of Sak K1ymazaslan (2008).

Discussion For RQ 2a And 2b

Intra-rater and inter-rater reliability are investigated through this research question. The results show the consistency of the scores given by a rater and between raters who worked as pair in the exam. In order to inspect intra-rater reliability Croncbach Alpha reliability analysis is carried out and for inspecting inter-rater reliability analysis Spearman correlation analysis which is the equivalent for Pearson correlation analysis for nonparametric test (Pallant, 2001) is carried out.

In the study carried out Sak K1ymazlaslan (2003), intra-rater reliability analysis is conducted by using Pearson correlation calculation between a rater's first and second ratings of the same performance which was recorded beforehand. According to the results, although

there is intra rater reliability for some raters, for some other raters intra rater reliability is not found. In this thesis study, on the other hand, intra rater reliability is calculated through Cronbach alpha analysis. Similar to the study by Sak K1ymazaslan (2003), while there is intra rater reliability for some raters, for some other raters there are not intra rater reliability. However, in this thesis study, it is found that there are highly more raters who have intra rater reliability than in the study by Sak K1ymazaslan. The reason for such a result could stem from different conditions. First of all, unlike this thesis study, in the study by Sak K1ymazaslan a rater training session was not held. Thus, inconsistency might have occurred since raters did not have a common view point on how to score and implement the exam. In addition to this, in the study by Sak K1ymazaslan (2003), raters had to rate the performances for the second time by listening to audio recordings. This might have also effected the perception of the raters about the performance of test takers.

In the study carried out Sak K1ymazlaslan (2003) and Zhou (2016), inter-rater reliability analysis is conducted by using Pearson correlation calculation between pairs of raters. According to the results for both of the studies, although there are inter rater reliability for some raters, for some other raters inter rater reliability is not found. In this thesis study, on the other hand, inter rater reliability is calculated through Spearman correlation. Similar to the study by Sak K1ymazaslan (2003) and Zhou (2016), while there are inter rater reliability for some raters, for some other raters there are not inter rater reliability. However, in this thesis study, it is found that there are highly more raters who have inter rater reliability than in the study by Sak K1ymazaslan (2003). The reason for such a result could stem from as stated above in the study by Sak K1ymazaslan (2003) a rater training session was not held. Also, similar to this thesis study, in the study by Zhou (2016) raters had rater training. However, when compared to this thesis study, in the study by Zhou (2016) there are less pairs with high inter rater reliability. This might result from different variables other than rater training, such

as usage of different types of assessment rubric like holistic rubric for one rater and analytic rubric for the other rater in the study of Zhou (2106). Different types of rubrics might have affected the way a rater assesses the performance. For instance, rater using analytic rubric might have taken a feature in the speaking performance into attention while the one who used holistic rubric might have loosely taken that feature into account and which could have caused low inter rater reliability.

Apart from the studies above, the other study conducted by Shatrova et.al. (2017) and in the examination of Checkpoint, rater training sessions were held. However, there are not statistical analysis of inter rater reliability in these studies. While in the study by Shatrova et al. (2017) raters who had low inter rater reliability were sent for revising the session again, in the examination of Checkpoint, raters who did not have acceptable rater reliability after rater training session, were not assigned in the examination. In addition to these examples, in the speaking examination held at University of Copenhagen for the lecturers, if there is a disagreement between the two raters, a third rater assesses the performance of the test taker and resolves the disagreement.

Discussion For RQ 3

In this thesis study, it is aimed to inspect the effectiveness and efficiency of the speaking exam in the eyes of participants who are administrators, teachers and students as test takers. Therefore, an interview is held by directing several questions to them.

In the study by Sak Kıymazaslan (2003), views of students are taken to see the efficiency of the speaking test. Authenticity of the exam is questioned through with the question with “To what extent did the speaking exam (...) reflect the characteristics of (...) in real life situations?” (p.160) which is asked to students and the question with “Are the tasks and topical contents relevant to the target language use domain ... the test taker is likely to

encounter?" (p.164) which is asked to teachers. Similar to these questions in study of Sak Kıymazaslan (2003), in this thesis study with the first question of the interview, the same answer is sought. In both studies, participants replied positively to the questions. It shows that the both tests have the potential of authenticity. Unlike this thesis study, in the study by Sak Kıymazaslan (2003), students' viewpoints are also taken to see the authenticity. This might have both advantages and disadvantages. For advantages, a different perspective could be added by asking this question to the students. On the other hand, since students may not be able to judge or foresee what kind of speaking contexts they might be in the future, the answers of them could be unrealistic. Therefore, in this thesis study, this type of questions are only asked to administrators and teachers.

In another interview question, in the study of Sak Kıymazaslan (2003), interviewees were asked how the test could be improved. With this question, whether further amendments are necessary is tried to be found out by the researcher. Similar to the study by Sak Kıymazaslan (2003), in this thesis study, with the second interview question same thing is aimed. According to the results of the question in the interview of Sak Kıymazaslan (2003), the two most selected answers are the inclusion of more interaction and the inclusion of various topics. Likewise, in this thesis study, some of the participants expressed that another dialogue task is necessary to be added to the test. However, in this thesis study, also one of the participants stated that need not to be removed but the type of explanation task should be renewed. In both of the studies, results are almost similar; however, in the study of Sak Kıymazaslan (2003) this question is asked to students while in this thesis study the questions are asked to both administrators and teachers.

In this thesis study, there are also some other questions that elicit answers regarding the practicality and washback effect of the exam. On the other hand, in the study of Sak Kıymazaslan (2003), the question asked to students elicit answers regarding their overall

content about the exam and questions asked to teachers are mainly related to content validity of the examination.



Chapter VI: Conclusion and Implications

Firstly, content validity of a speaking examination could be sustained by taking the field experts' view into account at the very beginning of developing a new test. In addition to the views of field experts, if the new developed test targets to test a certain proficiency level, aims and outcomes of the course book could also be taken into consideration. In the light of the views of field experts and aims and outcomes of the course book, test specifications should be defined and based on those specs tasks should be chosen or created. In order to investigate content validity of the test, abilities tested in the exam and aims and outcomes of the course book might be compared to find percentage of match. Also, interview could be held with teachers, administrators and even with students to inspect content validity of the exam.

Secondly, construct validity could be sustained by deciding on abilities to be tested, tasks, and assessment scale very carefully and suitably. As long as these three components of are developed by establishing a perfect alignment with each other, construct validity of an exam could be sustained mostly. Construct validity could be investigated in different statistical analyses such as through factor analysis, inter-item correlation analysis and multi-facet rasch analysis.

Thirdly, intra-rater and inter rater reliability could be sustained through a well programmed rater training session. In rater training, raters could be taught on how to assess properly and how to assign right score, differentiate between good, average and poor performances, executing the procedures to implement the speaking test and they should be given an opportunity of practice before the examination. The selection of raters who completes the training by maintaining intra and inter rater reliability would impact the consistency of the scores given by raters in the exam positively. Intra-rater reliability could be analyzed with Cronbach alpha reliability analysis or investigating the correlation between the

first and second score of a performance which must be recorded beforehand, and inter-rater reliability could be analyzed with the Pearson product moment correlation or Spearman correlation statistical analyses.

Finally, views of the participants who are administrators, teachers and students as test takers are important to be taken. By taking the views of those participants, it would be possible to evaluate the exam from different perspectives. In doing so, necessary amendments could be done to make the test better. In order to do that efficiently, questions to be asked should be decided carefully. Also, deciding on the which question to be asked which participant is also important since being academically qualified enough to state ideas on the aspects of an exam is crucial. It is possible to take the views of participants through interviews and questionnaires.

Pedagogical Implications

In the light of findings, discussion and conclusion parts, there are some pedagogical implications to be drawn. As stated above, validity and reliability issues hold great importance for a test. In this respect, there are some vital hints and tips to maintain validity and reliability dimensions of a test.

Content validity shows to what extent a test is representative of course book content, aims and outcomes. When content validity is established, test takers would be assessed more comprehensively and therefore results of the students would be a better indicator of what s/he is really capable of. Regarding the construct validity issue, Mislevy (2007) states that it is the base of any other type of validity. The reason might be that construct validity indicates whether a test is able to test what is intended. Therefore, blueprint of a test tasks and assessment scale ought to be aligned with each other. If the assessment scale is irrelevant of what type of speech the tasks in the exam are supposed to elicit, or vice versa, construct

validity could crack down. Thus, alignment among them is highly crucial. Also, when developing tasks, taking the views of field experts into account helps sustain and increase content validity and construct validity.

Reliability shows how much the scores are dependable and consistent. Without maintaining intra-rater reliability, establishment of inter-reliability could be open to debate. Also, without inter-rater reliability, score of a test-taker might not be a score that s/he might deserve. In order to establish both types of reliability, a thorough rater training is necessary. In this rater training, first of all raters should be instructed on the blueprint of the test, tasks, assessment scale and the procedures to implement the test. Next, the training should include sample performances of different levels and the distinctive features related to each level of performance. Then, a guided and semi guided practice on scoring of other different levels of performance on their own and raters having low level of inter-rater correlation might have extra theoretical guidance and extra exercise on scoring.

Finally, feedback is very important aspect of further improvements for the test. Views of participants in the process of whole speaking exam are very valuable for shedding light. Positive feedback would provide what the effective things are in the exam while negative one would show which parts of the exam could be more efficient. In this respect, directing questions through interview or questionnaire about the exam to administrators, teachers and test takers would make this possible. By choosing the questions carefully bearing in the mind the things that make a test perfect, invaluable feedback could be elicited from the participants.

Methodological Implications

One of the implications related to methodological aspect is about the analysis of construct validity of the test. Factor analysis is used for this purpose, because in the literature, most of the researchers mentioned in their articles about the usage of factor analysis for

investigating construct validity (Cronbach & Meehl, 1955; Fulcher, 2003; Hughes, 1989; Öztürk, 2007; Xiaoqing, 2003). However, MFRM is able to give out more detailed information on the interactions of a specific examiner with a certain examinee or a task and precision of the construct validity investigation would improve (Grabowski, K. C., 2009; Lynch & McNamara, 1998; McNamara & Roever, 2006; Rasch G. 1960). Therefore, an investigation of concurrent validity by using MFRM analysis would reveal more detailed results.

Another implication would be the analysis of concurrent validity of the test. Because of the size of the test takers, it would be impractical to have them sit another speaking examination of which validity is established. Therefore, concurrent validity is not investigated. However, with a careful sampling, the size of the test takers could be decreased and by granting a request to use online shared exams, the scores of the test takers in the exam and in the second exam could be investigated to find correlation.

Predictive validity of the speaking examination could be investigated by evaluating the test takers' performance in communication with tower and other pilots during flight training or communication with speakers of English as a foreign language in their international duties. However, since those contexts are not speaking examination contexts, correlation statistical analysis would be useless. Instead, an interview could be carried out with test takers in their future career to find how much effective the exam is in predicting the future success of a test taker in those contexts.

Raters only used analytic scale in this exam to make assessment. Although usage of analytic scales are fairer (Thornbury, 2005), Brookheart (2003) states that usage of holistic rubric provides high inter-rater reliability. Usage of holistic scales could make the assessment easier for raters since it requires the rater draw an overall impression rather than scoring by

focusing on many features at the same time. The consistency between the scores given by using holistic scale and the scores given by using analytic scores would not only provide more comprehensive assessment of the test taker by including another dimension but also would strengthen the reliability of the scores.

Finally, in this thesis study, views of participants are taken in order to see the efficiency of the examination. However, an important part of the exam is rater training as well. Views of participants could be taken to find the impact of rater training on their rating process. Views of participants on rater training would shed light on the efficiency of rater training and its role on scoring the performances of test takers.

Further Research

The effectiveness of rater training is an important research area to be conducted. An experimental study could be carried out to find the differences between the scores of raters having rater training and raters having no rater training. Also, by conducting an interview with them, more comprehensive results could be obtained and deeper insight could be gained.

References

- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.
- Andrews, F.M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 48, 409-442.
- Applebee, A. N. (2000). Alternative models of writing development. In R. Indrisano & J. R. Squire (Eds.), *Perspectives on writing: Research, theory, and practice* (pp. 90-110). Newark, DE: International Reading Association.
- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. California, CA: Corwin.
- Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*, 10 (2), 149-164.
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bailey, K. M., & Savage L., (1994). *New ways in teaching speaking*. Alexandria, VA:TESOL
- Berry, V. (1997). *Gender and personality as factors of interlocutor variability in oral performance tests*. Paper presented at the Language Testing Research Colloquium in Orlando, FL.
- Brookhart, S. M., & Nitko A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.
- Brown, H., D. (2004). *Language assessment: Principles and classroom practices*. New York, NY: Pearson Education.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saadle River, NJ: Prentice Hall Regents.

- Brown, J. D. (2005). *Testing in language programs*. New York, NY: McGraw-Hill.
- Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı* [Data analysis handbook for social sciences] (7th ed.). Ankara: Pegem A.
- Cambridge English. (2013, September 19). *Cambridge English: Preliminary for schools Victoria and Chiara* [Video File]. Retrieved from <https://www.youtube.com/watch?v=Qw1BZc9MrJA>
- CARLA of University of Minnesota. (2019). Task Specific Rubric. Retrieved from <http://carla.umn.edu/assessment/vac/improvement/rubrics/figure2.html>
- Cambridge English. (2014, April 23). *Cambridge English: Key speaking* [Video File]. Retrieved from https://www.youtube.com/watch?v=7D8_X5PzUpQ
- Cambridge English Language Assessment. (2016). Cambridge English preliminary handbook for teachers for exams from 2016. Retrieved from <https://www.cambridgeenglish.org/images/343147-cambridge-english-preliminary-for-schools-pet-for-schools-digital-handbook-for-teachers.pdf>
- Cambridge English Preliminary for Schools,(n.d.). *Handbook for teachers from 2016*. p.59. Retrieved from <https://www.cambridgeenglish.org/images/343147-cambridge-english-preliminary-for-schools-pet-for-schools-digital-handbook-for-teachers.pdf>
- Creswell, J.W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. New Jersey, NJ: Prentice-Hall, Inc.
- Cronbach, L. J., & Meehl, P. E. (1995). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343-368.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208-238.
- Fulcher, G. (2014). *Testing Second Language Speaking*. (2nd ed.). New York, NY: Routledge
- Fulcher, G., & Davidson F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge Publishing.
- Gay, L. R., & Airasian, P. (2003). *Educational research competencies for analysis and applications* (7th ed.). New Jersey, NJ: Prentice-Hall, Inc.
- Genesee, F., & Upshur, J.A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.
- Goh C.C.M., & Burns A. (2012). *Teaching speaking a holistic approach*. New York, NY: Cambridge University Press.
- Gong, B. (2010). *Considerations on conducting English speaking tests for advanced college students*. Retrieved from <https://docplayer.net/20576716-Considerations-of-conducting-spoken-english-tests-for-advanced-college-students-dr-byron-gong.html>
- Grabowski, K. C. (2009). Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking. PhD Thesis, Teachers College, Columbia University, New York, NY.
- Gronlund, N., & Linn, R. L. (1990). *Measurement and evaluation in teaching*. New York, NY: Macmillan Publishing Company.

- Hamill C. (2016). A mixed-methods approach to assessing the predictive validity of a language proficiency test for aviation training admissions.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.) *Assessing second language writing in academic contexts* (pp.241-276). Norwood, NJ: Ablex.
- Heaton, J. B. (1990). *Classroom testing*. Longman, London: New York, NY.
- Henning, G. (1987). *A guide to language testing: Development – evaluation – research*. Rowley, Massachusetts, MA: Newbury House.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge and New York, NY: Cambridge University Press.
- Hughes, A. (1990). *Testing for language teachers*. Glasgow: Cambridge University Press.
- Iwashita, N. (1999). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 8(1), 51-66.
- Kayi, H. (2006). Teaching speaking: Activities to promote speaking in a second language. *The Internet TESL Journal*, 12(11), Retrieved from <http://iteslj.org/Techniques/Kayi-TeachingSpeaking.html>
- Latitude Aviation English Services (2017). Checkpoint language testing for student selection test reliability and validity, Retrieved from <https://static1.squarespace.com/static/5742e4ef3c44d80a9c39295a/t/59115fa49de4bb22db4e2fec/1494310827155/Checkpoint+-+Test+reliability+and+validity+-+May+2017.pdf>
- Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective, *System*, 20, 373-386.
- Lazaraton, A. (2001). Teaching oral skills. In Marianne Celce-Murcia (Ed.) *Teaching English as a second foreign language* (pp. 103-113). Boston, MA: Heinle and Heinle.

- Lier, L.van (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23, 489-503.
- Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L.Odell (Eds.) *Evaluating writing: Describing, measuring, judging* (pp. 33-36). Urbana, IL: National Council of Teachers of English.
- Long, M. H., & Norris J. M. (2000). Task-based language teaching and assessment. In M.Byram (Ed.), *Encyclopedia of language teaching* (pp. 597-603). London: Routledge.
- Lu, Zhihong & Li, Zhenxiao & Hou, Leijuan. (2016). On the Validity and Reliability of a Computer-assisted English Speaking Test. *International Conference on Intelligent Control and Computer Application* (pp. 187-193). Retrieved from <https://pdfs.semanticscholar.org/52ed/32e42a4337674fd79bb469e621916f7adbbb.pdf>
- Lynch, B., & McNamara, T. (1998). Using G-theory and many-facet rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing*, 8, 14-33.
- McMillan, J. H., & Scumacher, S. (2006). *Research in education, evidence-based enquiry*. (6th ed.).Boston, MA: Allyn and Bacon.
- Mcnamara, T. (1996). *Measuring second language performance*. London: Longman.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Maiden, MA: Blackwell.
- Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10(9), 9-20.

- Messick, S. (1989). Meaning and values in test validation, the science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.
- Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, 36(8), 463–469.
- Moskal, B. M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment Research & Evaluation*, 7(3). Retrieved from <http://PAREonline.net/getvn.asp?v=7&n=3>.
- Nakamura, Y. (2014). Theoretical and practical issues of assessing speaking in university entrance examinations. *Educational Studies*. 56, 175-180.
- Norton, J. (2005). The paired format in the Cambridge speaking test, *ELT Journal*, 59(4), 287-297.
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: CUP.
- O'Malley J. M., & Pierce V. L. (1996). *Authentic assessment for English language learner: Practical approaches for teachers*. Reading, MA: Addison Wesley Publishing Comp.
- O'Sullivan, B. (2002). Learner acquaintanceship and OPT pair-task performance. *Language Testing*, 19, 277-295.
- Pallant, J. (2001). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Versions 10 and 11)*. Philadelphia, PA: Open University Press.
- Plakans, L. (2017). *A video on integrated assessment*. Retrieved from <http://languagetesting.info/video/main.html>
- Petersen, W. (1999). *50 French oral communication activities with mini-rubrics*. Auburn Hills, MI: Teacher's Discovery. Retrieved from: <http://carla.umn.edu/assessment/vac/improvement/rubrics/figure7.html>

- Popham, W.J. (2003). *Test better, teach better: The instructional role of assessment* (3rd ed.). Alexandria: ASCD.
- Qin Xiaoqing. (2003). *Quantitative statistic analysis in foreign language teaching research*. Wuhan, China: Central China Science and Technology University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: Danish Institute for Educational Research.
- Razi, S. (2012). Developing the inventory of cultural components to assess perception in language learning, *Novitas-Royal* (Research on Youth and Language), 6(2), 169-186.
- Reves, T. (1991). From testing research to educational policy: A comprehensive test of oral proficiency. In J. C. Alderson and B. North (Eds), *Language testing in the 1990s*. London: Modern English Publications and the British Council, (pp. 178-188).
- Sak Kıymazaslan G., (2008). *An investigation of the validity and reliability of the speaking exam at a Turkish university* (Master's Thesis, METU, Ankara). doi=10.1.1.632.5852
- Savignon, S. (1985). Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. *The Modern Language Journal*, 69, 129-134.
- Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal*, 40, 212-220.
- Standard Error Calculator. (2019, May 05). Retrieved from: <https://www.socscistatistics.com/tests/standarderror/default.aspx>
- Stirling, B. (2012). *Practice Tests for TOEFL IBT*. Nova Press.
- Strauss, A. L., & Corbin, J. (1997). *Grounded theory in practice*. Thousand Oaks, CA: Sage.
- Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18(3), 275-302.
- Tedick, D. J. (2002). Proficiency-oriented language instruction and assessment: Standards, philosophies and considerations for assessment. In D. J. Tedick (Ed.) *Proficiency-*

- oriented language instruction and assessment: A curriculum handbook for teachers* (Rev Ed.). CARLA Working Paper Series, Minneapolis, MN: University of Minnesota. Retrieved from: https://carla.umn.edu/articulation/polia/polia_intro.pdf
- Thornbury, S. (2005). *How to teach speaking*. UK: Pearson.
- UCLES, (2014). *Cambridge English assessing writing performance – level B1*. Retrieved from <https://www.cambridgeenglish.org/images/231794-cambridge-english-assessing-writing-performance-at-level-b1.pdf>
- UCLES, (2016). *Cambridge English preliminary handbook for teachers*. Retrieved 2019 from <https://www.cambridgeenglish.org/Images/168150-cambridge-english-preliminary-teachers-handbook.pdf>
- Underhill, N.(1987). *Testing spoken language*. Cambridge: Cambridge University Press.
- Ur, P. (2000). *A course in language teaching: Practice and theory*. Cambridge: Cambridge University Press.
- Wei, L. (2011). Validity considerations in designing and oral test. *Journal of Language Teaching and Research*, 2, 267.
- Weir, C. (1993). *Understanding and developing language tests*. New York, NY: Prentice Hall.
- Wesche, M. (1987). Second language performance testing: the Ontario test of ESL as an example, *Language Testing*, 4, 28-47.
- Zhou, W., (2016), Investigating the construct validity of communicative proficiency in TEP (Oral) at Level B, *Journal of Language Teaching and Research*, 7(4), 690-699.

Appendices

Appendix A: Scoring Sheet For Raters

____ DEKANLIĞI
 ____ EÖY HAZIRLIK SINIFLARI
 İNGİLİZCE B1 SINAVI
 KONUŞMA SINAVI

Sınav Saati			Derslik / Komisyon		TASK 1							TASK 2					TOPLAM	
					Topic Explanation	Sınav Görevlisi Değerlendirme Notu					Interaction	Sınav Görevlisi Değerlendirme Notu						
S/N	NO	SINIFI	ADI	SOYADI	TASK NO	G	V	P	F	C	T	TASK NO	G	P	V	IC		T
1																		
2																		
3																		

Sınav Görevlisi Ad Soyad: İmza

Appendix B: Scoring Sheet for Taking Average Score

_____ DEKANLIĐI
____ EÖY HAZIRLIK SINIFLARI
İNGİLİZCE B1 SINAVI
KONUŞMA SINAVI KOMİSYON _____

S/N	NO	SINIFI	ADI	SOYADI	1'inci Sınav Görevlisi Değerlendirme Notu	2'nci Sınav Görevlisi Değerlendirme Notu	Ortalama Not
1							
2							
3							
4							
5							

Appendix C: Sample Question for Sustained Monologue Production Task

1. Friendship is the most important relationship; do you agree or disagree? Why? How do you maintain a good friendship? What things should friends never do?

- If your friendship gets broken, what would you do to fix the relationship?

2. What is the importance of doing sport? What kind of sports do you do? Why? What other kinds of sports would you like to learn? Explain your ideas by giving reasons and examples.

- Do you agree or disagree with the idea that “athletes earn much more than they deserve? Why? Why not?

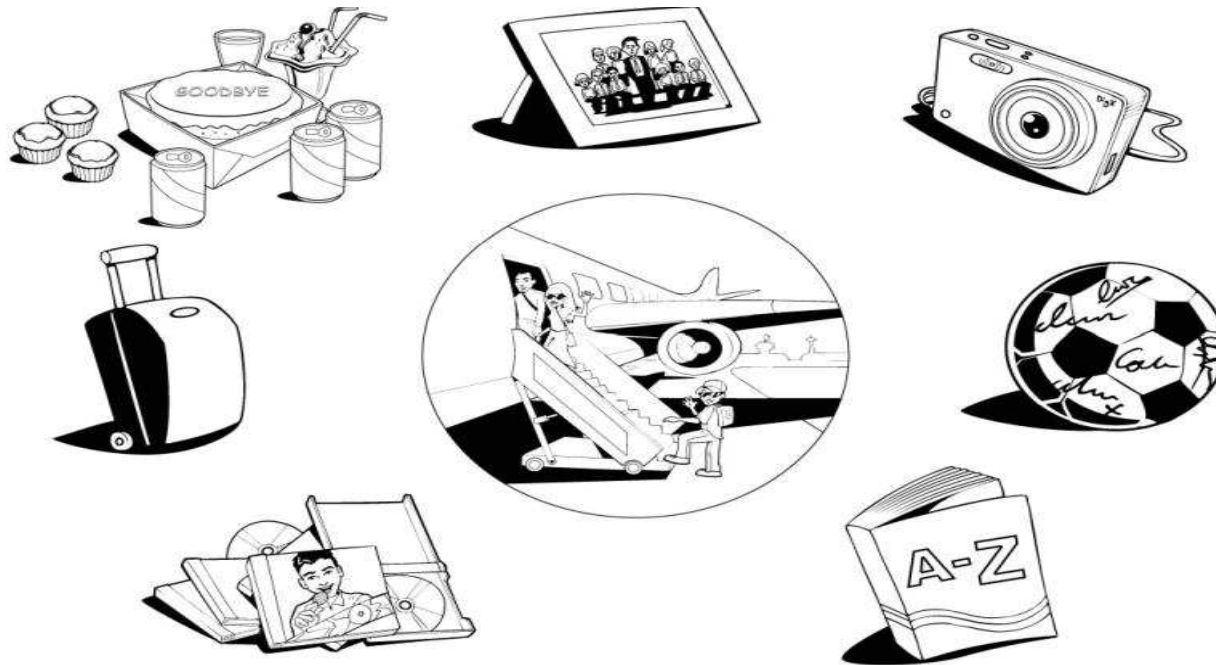
Appendix D: Sample Question for Pair Interaction Task (Cambridge English Preliminary for Schools, n.d, p.59)

TASK 2 TASK NUMBER 1: A boy is leaving his school because his parents are going to work in another country. The students in his class want to give him a present.

*Ask questions to each other about the each present below.

* Answer each other's questions by giving reasons.

* Come up with a decision together about what they can give as a present



Appendix E: Assessment Rubric for Task 1 (Sustained Monologue Production)

TASK 1 (Sustained Monologue Production)	10-9	8-7	6-5	4-3	2-0
Grammar	Shows a good degree of control simple grammatical forms and attempts some complex grammatical structures.	Performance features share Band 10-9 and Band 6-5	Shows sufficient control and simple grammatical forms.	Performance features share Band 6-5 and Band 2-1	Shows only a limited control of a few grammatical forms.
Vocabulary	Uses a range of appropriate vocabulary to give and exchange views on familiar topics.	Performance features share Band 10-9 and Band 6-5	Uses sufficient and appropriate vocabulary when talking about familiar topics.	Performance features share Band 6-5 and Band 2-1	Uses insufficient and inappropriate vocabulary to talk about familiar topics.
Pronunciation	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Performance features share Band 10-9 and Band 6-5	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Performance features share Band 6-5 and Band 2-1	Is mostly intelligible, despite limited control of phonological features.
Fluency	Spontaneity of expression, relatively ease of expression even in longer stretches. Few pauses for search of grammatical and lexical planning	Performance features share Band 10-9 and Band 6-5	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident	Performance features share Band 6-5 and Band 2-1	Serial short periods of silence. Several gaps that disrupt the flow of information. Listeners' attentions diverted to the gaps rather than the message.
Production	Produces extended stretches of language despite some hesitation. Contributions are relevant despite some repetition. Uses a range of cohesive devices.	Performance features share Band 10-9 and Band 6-5.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, but there may be some repetition. Uses basic cohesive devices.	Performance features share Band 6-5 and Band 2-1	Produces responses which are characterized by short phrases and frequent hesitation. Repeats information or digresses from the topic.

Appendix F: Assessment Rubric for Task 2 (Pair Interaction)

TASK 2 (Pair Interaction)	10-9	8-7	6-5	4-3	2-1
Grammar	Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms	Performance features share Band 10-9 and Band 6-5	Shows a good degree of control of simple grammatical forms.	Performance features share Band 6-5 and Band 2-1.	Shows sufficient control of simple grammatical forms.
Pronunciation	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Performance features share Band 10-9 and Band 6-5-	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Performance features share Band 6-5 and Band 2-1.	Is mostly intelligible, despite limited control of phonological features.
Vocabulary	Uses a range of appropriate vocabulary to give and exchange views on familiar topics.	Performance features share Band 10-9 and Band 6-5-	Uses a range of appropriate vocabulary when talking about familiar topics.	Performance features share Band 6-5 and Band 2-1.	Uses a limited range of appropriate vocabulary to talk about familiar topics.
	20-17	16-15	14-11	10-8	7-3
Interactive Communication	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.	Performance features share Band 20-17 and Band 14-11.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.	Performance features share Band 10-8 and Band 7-3.	Maintains simple exchanges, despite some difficulty. Requires prompting and support.

Appendix G: Official Permission Form for Using Online Sources of Cambridge English Language Assessment

Cambridge University Press

Sayı: 001/05.2019.....

Konu: Online Materyal Kullanım İzni

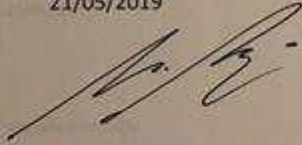
Sayın Vakkas Selim YÜKSEL

Yabancı Diller Bölüm Başkanlığı

Yürütmekte olduğunuz "Developing a Valid and a Reliable Speaking Test in the Frame of Institutional Requirements" konulu yüksek lisans tezinizde yararlanmak istediğiniz Cambridge English Language Assessment tarafından hazırlanan bütün çevrimiçi kaynaklarımızı kullanmanızda Kurumumuz tarafından herhangi bir sakınca bulunmamaktadır.

Rica ederim.

21/05/2019



Mustafa Akıncioğlu

Research, Product Development, Higher Education Manager

Cambridge University Press

İçerenköy Mahallesi,

Üsküdar-İçerenköy Cad. No:10, Kat:11

34752, Ataşehir / İstanbul

Appendix H: Official Permission Form for Using Online Sources of The Center for Advanced Research on Language Acquisition (CARLA) of University of Minnesota

https://mail.google.com/mail/u/0/?ui=2&view=bt&ver=1mqabpd1bb0qt&search=inbox&th=%23thread-f%3A1640435005309505307&cvid=1

Copyright Permission Granted: CARLA Website materials Gelen Kutusu x

Karin Larson <larso205@umn.edu> 30 Temmuz Sal 01:52 (5 gün önce) ☆ ↶ ⋮
 Alıcı: ben, syuksel ▾

İngilizce ▾ > Türkçe ▾ İletiyi çevir İngilizce için kapat x

Dear Vakkas Selim YÜKSEL:

CARLA is pleased to give you permission to use the materials from the CARLA website for your dissertation. Citations of these materials should be listed as follows (or similar):

Figure 2. Task-specific rubric for a presentational writing task: "Visiting monuments in Paris" - intermediate level. (n.d.). In the Center for Advanced Research on Language Acquisition's Virtual Assessment Center. Retrieved from <http://carla.umn.edu/assessment/vac/improvement/rubrics/figure2.html>.

Tedick, D.J. (2002). Proficiency-oriented language instruction and assessment: Standards, philosophies, and considerations for assessment. In D.J. Tedick (Ed.), *Proficiency-oriented language instruction and assessment* (pp. 9-47). Minneapolis: University of Minnesota, The Center for Advanced Research on Language Acquisition. Available from https://carla.umn.edu/articulation/polia/polia_intro.pdf.

Task and multiple trait mini-rubric. (n.d.). In the Center for Advanced Research on Language Acquisition's Virtual Assessment Center. Retrieved from <http://carla.umn.edu/assessment/vac/improvement/rubrics/figure7.html>.

With best wishes,

Karin Larson
 --

Karin E. Larson
 Executive Assistant Director
 The Center for Advanced Research on Language Acquisition (CARLA)
 Global Programs and Strategy Alliance
 University of Minnesota
 140 University International Center
 324 47th Avenue Southeast

Appendix I: Official Permission Form for Using *Q:Skills for Success Listening and Speaking 3* Course book of Oxford University Press

PERMISSIONS LETTER

Our Reference: 4056020

Vakkas Selim YÜKSEL
National Defence University Turkish Air Force Academy
Turkish Air Force Academy / Yeşilyurt-İSTANBUL
34149 Turkey

16/07/2019

Dear Mr. YÜKSEL

Re: *Q:Skills for Success Listening and Speaking 3* by Miles Craven, Kristen D. Sherman, Marguerite Ann Snow, Lawrence J. Zwier, Cherly Boyd Zimmerman ISBN: 9780194056020
The Material: Scope and sequence, xiv, xvi, xviii

Oxford University Press (OUP) is pleased to grant you, a Student of National Defense University Turkish Air Force Academy, non-exclusive permission free of charge to use the material described above, subject to the following terms and conditions:

- 1 Use of the Material shall be restricted to inclusion in a dissertation by Vakkas Selim YÜKSEL in the English language entitled *Developing a Valid and a Reliable Speaking Test in the Frame of Institutional Requirements* which shall not be published or sold.
- 2 This permission shall be limited to the particular use authorized in 1 above and does not allow you to sanction its use elsewhere.
- 3 The Material shall not be altered in any way without our written agreement.
- 4 If the Material includes licensed content such as extracts, papers or illustrations reproduced from other publications or sources and where it is indicated in an Acknowledgements list or in any other manner in the Work that permission to use or include such material is required then permission must be sought from the copyright owner to cover the use of such material and to pay the copyright owner any necessary reproduction fees.
- 5 The following credit line appears wherever the Material is used or is included in an acknowledgement list either at the beginning or at the end of the survey and dissertation:


Reproduced by permission of Oxford University Press
from *Q:Skills for Success Listening and Speaking 3* by Miles Craven, Kristen D. Sherman, Marguerite Ann Snow, Lawrence J. Zwier, Cherly Boyd Zimmerman © Oxford University Press 2011

Please return one signed copy of this letter to Erika Lastovskyte, ELT Partnerships and Innovation, to indicate your acceptance. Until ELT Partnerships and Innovation receives a signed copy of this letter, use of the Material is unauthorized.

Yours sincerely


Erika Lastovskyte (Jul 16, 2019)
Erika Lastovskyte
ELT Partnerships and Innovation

Accepted and Signed by


Vakkas Selim YÜKSEL (Jul 16, 2019)
For National Defence University Turkish Air
Force Academy